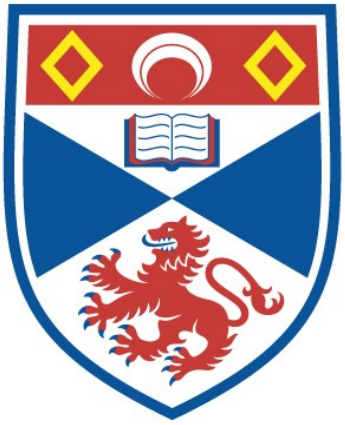
ON THE USE OF GENERATING FUNCTIONS FOR TOPICS IN CLUSTERED NETWORKS

Peter Mann

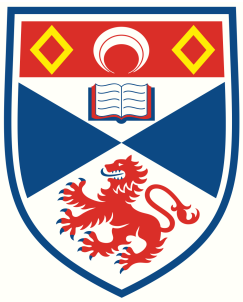
A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2022

On the use of generating functions for topics in clustered networks

Peter Mann



University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy

at the University of St Andrews

September 2021

Dedicated to my Wife, Daughter, Mum and Dad.
Thank you for all the love you bring.

Candidate's declaration

I, Peter Stephen Mann, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 80,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.
I was admitted as a research student at the University of St Andrews in March 2017.
I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date 17Feb2022

Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 17Feb2022

Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.
I, Peter Stephen Mann, confirm that my thesis does not contain any third-party material that requires copyright clearance.
The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy
No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date
17Feb2022

Signature of candidate

Date
17Feb2022

Signature of supervisor

Underpinning Research Data or Digital Outputs
Candidate's declaration
I, Peter Stephen Mann, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date
17Feb2022

Signature of candidate

Acknowledgements

I would like to thank my family for their love and support throughout this PhD. This qualification is the culmination of my time at St Andrews, and my prior education, and I would like thank my parents for their sacrifices that enabled me towards this goal. As a father, I hope to continue this commitment to education for my daughter, Hallie.

I owe a great deal to Professor Simon Dobson, who's guidance, mentoring, intellect and friendship has been truly invaluable during this journey. I feel very privileged to be his student. I would also like to thank Dr John Mitchell and Dr Anne Smith for their time, fierce intellect and guidance.

I would like to thank my wife, Kirsten, for everything that she does and for being my best friend.

This work was supported by the University of St Andrews (School of Chemistry and the School of Biology).

Kind regards,

Peter Mann

Abstract

In this thesis we relax the locally tree-like assumption of configuration model random networks to examine the properties of clustering, and the effects thereof, on bond percolation. We introduce an algorithmic enumeration method to evaluate the probability that a vertex remains unattached to the giant connected component during percolation. The properties of the non-giant, finite components of clustered networks are also examined, along with the degree correlations between subgraphs. In a second avenue of research, we investigate the role of clustering on 2-strain epidemic processes under various disease interaction schedules. We then examine an N-generation epidemic by performing repeated percolation events.

PRIOR PUBLICATION

This thesis has led to the publication of 6 first-author papers in Physical Review E [32, 33, 31, 35, 36, 34]. All of the papers concern some aspect of the role of clustering in random graphs. The content of both [32] and [33] is largely not discussed in this thesis; since, they are approximate methods. Instead, I chose to focus on exact solutions given by the other papers. There is also a fair amount of unpublished material including chapters 4, 6 and 8 and section 3.4.

CONTENTS

CHAPTER ONE

NETWORKS AND THEIR COMPLEXITY

A complex system is a system that is composed of many individual parts that display collective, global behaviour that does not follow trivially from considering the microscopic. A complex system is more than the sum of its parts. Due to this abstract definition, complexity can be observed in almost all areas of natural and computational science; it is a broad field transcending traditional boundaries, that encompasses a wide range of analytical and experimental methods. Network science has an almost innumerable list of applications including condensed matter systems, ecosystems, economies, markets, biochemical networks, and certainly, human contact networks fall under the umbrella of complexity.

Network science is the unification of topology and dynamics; whom interacts with whom, as well as the nature of the interaction itself. The scaling and criticality (phase behaviour) of the dynamics is dependent on the nature of the contact topology; however, interestingly, this is often independent of the details of the system under study. For instance, swarm behaviour observed in schools of marine life is semantically similar to the emergent patterns that occur when birds group together during their mating season. The global magnetisation that results from spin alignment within an individual ferromagnetic crystal domain at the Curie temperature is also broadly similar to an avalanche that results from a sudden collective behaviour of many snowflakes.

A mathematical model is a representation of reality. For instance, consider the timely example of a disease spreading through a population. Epidemiological knowledge tells us that there must be some vector of transmission from an infected person to a susceptible person. A traditional modelling procedure is to assume, in the first instance, that the population is well mixed; in other words, we ignore the role of contact structure and concentrate on the dynamics. This first approximation might be well justified; however, it may also be overlooking a fundamental aspect of the reality. Perhaps, the contact structure plays a significant role in the outbreak pattern in the population. Indeed, the well mixed model assumes that an airborne flu or common cold spreads in a similar fashion to a sexually transmitted disease such as HIV. Network science allows us to model, and therefore to understand, the role that contact structure plays in the dynamics. It enables topology to be another variable in the mathematical model; and often, we see significant variation in the dynamical properties as a result.

1.1 Network definitions

To discuss network structure, we must understand some metrics from graph theory. A network is a collection of individuals, which we call *vertices*, that are connected together by links, which we call *edges*. To a mathematician a network is a *graph*, and we use the two terms indiscriminately. The number of edges that a particular vertex has is called its *degree*. For instance, in a social network of human interactions, a person with lots of contacts has a high degree, whilst a person with few contacts has a low degree. The probability distribution of choosing a vertex from the network that has degree k is the *degree distribution*, p_k. The degree distribution is a fundamental descriptive object for a network as many other properties can be calculated from it.

There are many different kinds of networks that are classified by some distinguishing property. The property could be the colour of vertices or edges when arranged into layered networks, or it might be the presence or absence of closed loops in clustered networks, or it might be the analytical form of the degree distribution.

A *random graph* is a graph that has been created by randomly connecting a collection of degree-labelled vertices together according to some stochastic algorithm. Many different random graphs can be made by repeating the construction process and thus a given random graph belongs to an ensemble of such networks. This is in contrast to an *empirical network*, which is a particular realisation of a graph that has been collected from some observation in a field of study. Any given graph can be mined for its properties such as degree distribution, its number of closed loops, its average path length or diameter. It is often the case that analytical methods are well equipped to describe the mean properties of the ensemble of random graphs, but are not so equipped to single out a particular realisation. Therefore, the properties of real-world empirical networks often become inferred from the mean field description of random graph ensembles that can be analytically studied by a mathematical model.

A model might start by assuming that all of the contacts are tree-like [7, 54], that is to say, that there are no short cycles or loops among the vertices, see Fig 1.1. This removes correlations (feedback circuits) between a vertex's edges and makes the analytical model quite straightforward. Models based on locally tree-like approximations have provided much insight into the properties of random ensembles and empirical networks [37]. A tree-like approximation might well be a valid representation of a given empirical network; however, it is unlikely that a social network of human contacts would be tree-like. This would suggest that an individual's friends did not know one another, which is unrealistic. It is natural to ask how might the induced correlations brought about by the presence of short-range loops in the contact structure effect the structural properties of the network? In turn, how do these structural changes effect dynamical processes that occur over the network? For instance, does the clustering of social contacts allow an epidemic to spread more easily through a population or does it make it more difficult?

To examine the effects of loops in random networks we must relax the locally tree-like assumption. Such networks are known as *clustered networks*; and have received much attention in the literature using a variety of different analytical approaches [39, 28, 52, 20, 22, 18, 47, 11, 5, 63, 62, 32, 19]. From this extensive work in the clustered-graph literature, it has become apparent that not only does the presence of short-range loops bear significant effect on the properties of networks, but also *how* those loops are structured with respect to one another is important. Not all clustered networks are equivalent. Consequently, there are many different ways to induce clustering into a random graph model.

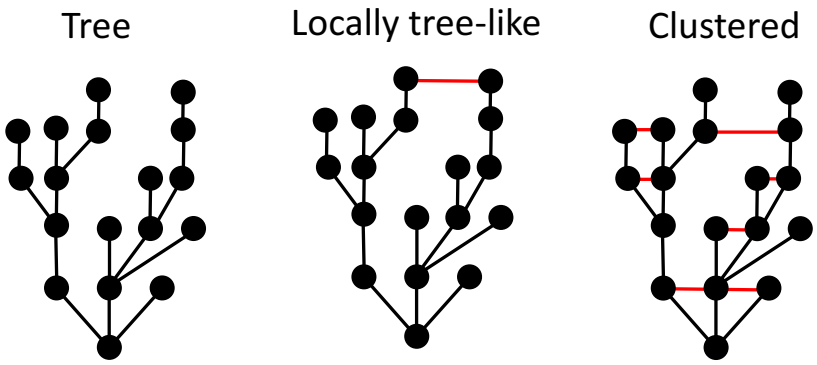


Figure 1.1: The presence of loops (red) among vertices causes correlation among the edges of the network. When the network is locally tree-like, analytical methods based on trees provide good approximations [37]. For clustered networks these approximations often fail to describe the details of dynamical processes [47, 40, 52]. There is no strict definition of when a locally tree-like network is a clustered network.

1.2 Structure of this thesis

In this thesis we will examine the properties of clustered networks. We use the *configuration model* (see section 1.5) as a means to create ensembles of random graphs for simulation purposes and the *generating function formulation* as an analytical toolkit to study their properties theoretically. We are mostly concerned with the *bond percolation process*, described in section 1.3, which also has strong links to the SIR dynamical process from epidemiology, see section 1.4.

There are two themes in this research: i) the analytical treatment of clustered networks; ii) the study of epidemic processes on clustered networks. Chapters 3, 4, 5 and 6 fall under the first branch of this research, whilst chapters 7 and 8 examine the latter.

1.3 Bond percolation: the GCC and the RG

Bond percolation is a stochastic process developed by Broadbent and Hammersley in 1957 and is perhaps one of the simplest models that exhibits complexity. It examines the consequences, to the global connectivity of a network, of adding or removing connections between each possible pair of vertices. For a network comprising an unconnected set of vertices, bond percolation attempts to connect adjacent vertex pairs with a fixed and statistically independent probability T. In the special case that $T = 0.0$, no connections are added to the network; conversely when $T = 1.0$, all possible connections are created. Equivalently, bond percolation can also be conducted over the edges of a substrate network. In this case, edges are *occupied* with probability T and *unoccupied* with probability $1 - T$. Once all of the edges of the substrate network have been examined in this way, the process has reached its absorbing state. This equilibrium is static in nature; since, there are no more

dynamics to occur. As experiments are conducted at larger T values within the unit interval, the absorbing state network becomes increasingly connected by occupied edges. At some critical value, T_c, known as the *percolation threshold*, small clusters of occupied edges join up and large clusters of connected vertices span the graph. The largest of these clusters is known as the *giant connected component* (GCC). Vertices that do not belong to the GCC belong instead to the *residual graph* (RG), see Fig 1.2. The percolation threshold marks the value of T at which the GCC first appears and long range connectivity is established.

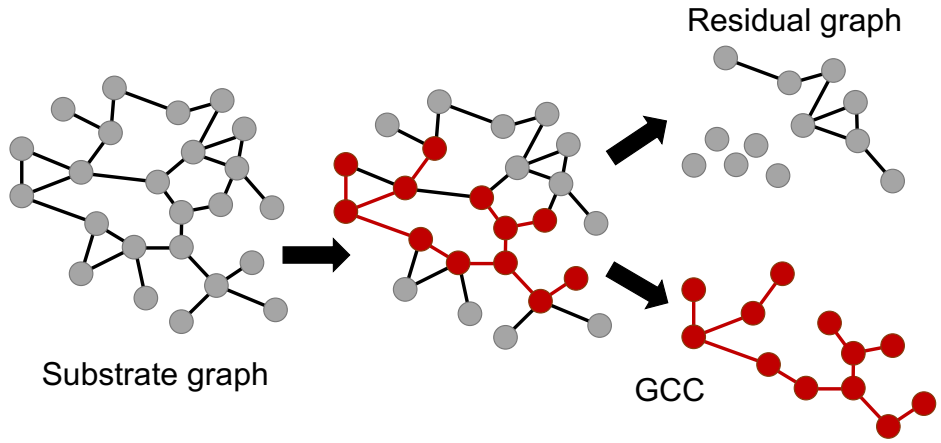


Figure 1.2: (left to right) A substrate network has undergone bond percolation to create a GCC of vertices connected by occupied edges (red) and a RG of vertices not contained within the GCC (grey).

Bond percolation is concerned with the global connectivity of a network, which depends on the sizes of the connected clusters. However, the presence of global connectivity also depends on the size of the underlying network; a cluster of 200 connected vertices will span a small network, yet is barely visible within a very large network. To address this formally, we must replace the finite size of the network with an infinitely large system. The central question of bond percolation now concerns the presence of an infinitely spanning cluster. By Kolmogorov's zero–one law, for any given T, the probability that an infinite cluster exists is either zero or one. The zero-one law indicates that the properties of the infinite system cannot be altered by finite perturbations. For instance, the presence of an infinite cluster in an infinite network is undisturbed by the addition or removal of a finite number of edges; which is not true for a finite sized network. For a given value of T, an infinite cluster will exist with probability one or zero; when $T < T_c$ the probability is zero, whilst for $T \geq T_c$ the probability is one.

It is Kolmogrov's law that leads to the phase behaviour observed in infinite systems. Below the percolation threshold, the probability that there is an infinite spanning cluster is zero. We could be forgiven for thinking that as we infinitesimally increase T to $T + \varepsilon$ where ε is a very small constant, that the probability of finding an infinite cluster might also increase in a gradual fashion. However, by the zero-one law, the infinite cluster is either present or not. At and above the percolation threshold, we are guaranteed to find a spanning cluster in an infinite system if we examine it close enough. This behaviour leads to a *phase transition* in the global connectivity of the infinite system, see Fig 1.3.

The presence of the GCC follows the zero-one law; however, its *size* does increase in a gradual fashion with increasing T. Finding the size of the GCC is a central topic

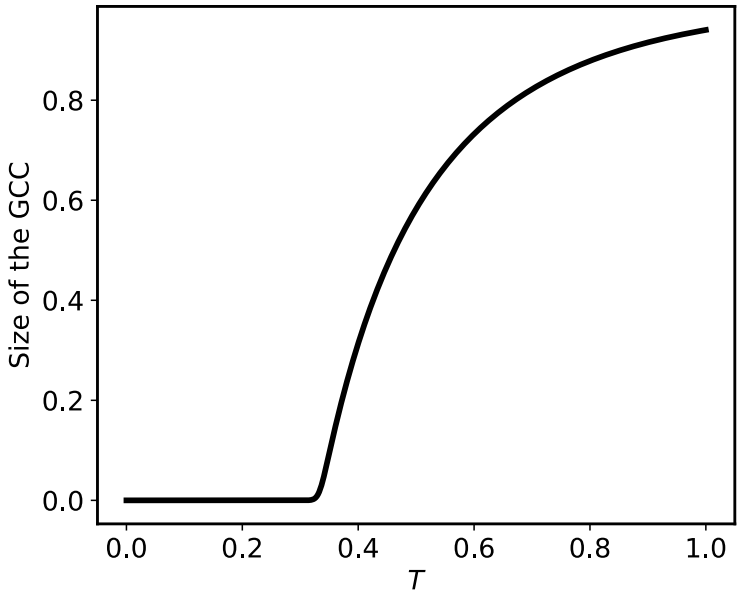


Figure 1.3: The size of the GCC following bond percolation on Erdős-Renyi networks with $\langle k \rangle = 3$ as a function of $T \in [0,1]$. Below the critical point $T_c = 1/3$ the size of the GCC is zero, whilst at and above T_c the GCC occupies a finite fraction of the network, growing at different rates as a function of T.

for network science and this thesis; it is the *order parameter* of the percolation problem. The location of the critical percolation probability and the rate of growth of the GCC are extremely dependent on the topological properties of the substrate network. For instance, the critical point of Erdős-Renyi networks occurs at $T_c = 1/\langle k \rangle$ where $\langle k \rangle$ is the average degree of the network. If the vertices of the substrate network have a larger number of edges on average, then the probability of forming a GCC increases and so the threshold value of bond occupancy probability T_c drops. Conversely, the critical point in scale-free networks occurs at $T_c = 0$ and hence, a GCC can always be found.

1.4 Epidemics on networks

The study of the spread of disease on networks has attracted considerable attention from the statistical physics community [45]. This unusual fact is a consequence of the equivalence between the properties of the susceptible-infected-removed model (SIR) and those of the bond percolation process, which has an exact solution on networks. The SIR model is an epidemiological model that considers the a population to be divided into three states: susceptible (S), infected (I), and removed (R). Infected individuals can pass their infection across an edge to a susceptible neighbour, which in turn becomes infected. The average infection rate per unit time is β. At each moment, the infected vertex might recover with a given recovery rate per unit time, γ, and no longer play an active part in the dynamics of the process. The equilibrium of the process is reached when all of the infected vertices have recovered and it is a static absorbing state comprising of recovered vertices and susceptible

vertices that did not become infected. This is the key part of the equivalence to bond percolation: the binary nature of the equilibrium.

Consider a particular infected vertex i and its susceptible neighbour j. We assume that the infectious period for all vertices is fixed to τ units of time. The probability that the edge fails to transmit the infection is $1 - T_{ij}$, which is

$$1 - T_{ij} = 1 - (1 - \beta_{ij})^{\tau} \tag{1.1}$$

where β_{ij} is the probability per unit time that the edge transmits infection which is an independent and identically distributed (iid) random variable. The average transmission probability, T, is found by integrating over β [21, 45, 66, 61]

$$T = 1 - \int_0^{\infty} P(\beta)(1-\beta)^{\tau} d\beta \tag{1.2}$$

which takes values in the unit interval. Thus despite variation in the individual transmission probabilities between different vertex pairs, the disease propagates as if all transmission probabilities were equal to T. This is precisely the ordinary bond percolation model. Therefore, the minimum transmissibility that a disease must have in order to create a population-wide outbreak is given by the percolation threshold; similarly, the size of the outbreak is equivalent to the size of the GCC.

We remark that the infectious period is assumed to be drawn from a single-valued distribution for all vertices. If we relax this assumption, then the mapping between percolation and SIR no longer holds; but, it can be recovered by considering percolation on a semi-directed networks instead [38, 27].

1.5 The configuration model

The configuration model is a protocol to create random graphs from a given sequence of vertex degrees. Upon each construction, a particular random graph is obtained from an ensemble of degree-equivalent, uncorrelated random graphs. It might happen that the realisation is not well represented by the ensemble average; therefore, simulations of dynamical processes over these networks require repetitions to establish the mean field.

In the model, the vertices of the graph are assigned an integer, drawn at random from the degree distribution, which indicates its degree. The degree sequence $\{k\} = k_1, k_2, \ldots, k_V$, where $\sum_i k_i = 2E$ for a network of $V \in \mathbb{N}$ vertices and $E \in \mathbb{N}$ edges, is a sequence of the degrees of the vertices and is typically displayed in descending order such that $k_1 \geq k_2 \geq \cdots \geq k_V$. To construct configuration model networks from a degree sequence, vertex i is inserted k_i times into a set for all $i \in V$. Pairs of vertices are then drawn at random and connected together. In the limit of large and sparse networks, the probability that the construction process chooses pairs that are either already connected through another edge or belong to the same vertex is vanishingly small. These networks are said to be *locally tree-like*, meaning that cycles have zero measure when $V \to \infty$ and hence the networks contain no short-range loops; they are also absent of degree-correlations.

It is often the case that a given degree sequence can be used to construct an ensemble of different graphs. However, not all degree sequences are valid, or *graphic*, such that some sequences of integers cannot be used to create a graph. The Erdős-Gallai theorem (EGT) states that in addition to the handshaking lemma (HL), $\sum_i k_i = 2E$, a sequence is

graphic if and only if the Erdős-Gallai inequality (EGI)

$$\sum_{i=1}^{l} k_i \leq l(l-1) + \sum_{i=l+1}^{V} \min(k_i, l) \tag{1.3}$$

holds for $1 \leq l \leq V$. It is trivial to construct degree sequences that satisfy the HL (EGI) but do not satisfy the EGI (HL) and are thus not graphic. For instance, with $V = 3$ and $\{k\} = \{(1),(1),(1)\}$ the inequality in Eq 1.3 is satisfied but the sum of degrees is not even; whilst, $\{k\} = \{(2),(0),(0)\}$ satisfies the lemma but not Eq 1.3.

1.6 The generalised configuration model

The generalised configuration model (GCM) is a protocol to create random graphs whose vertices can be members of predefined subgraphs. Therefore, it allows us to construct random graphs with tailored clustering that we can fine tune by allowing different subgraphs to be part of the model, in different quantities. In the GCM, the degree distribution is replaced by a joint degree distribution that describes a vertex's involvement in motifs such as ordinary edges (2-cliques), triangles (3-cliques), 4-cliques etc. For instance, consider a GCM network with 2-cliques and 3-cliques, a vertex that is involved in n_2 ordinary edges and n_3 triangles is specified by joint degree (n_2, n_3) and the usual degree is recovered from $k = n_2 + 2n_3$. Similarly, allowing 4-cliques into the topology set, the joint degree of a vertex that is a part of n_2 ordinary edges, n_3 triangles and n_4 4-cliques is given by (n_2, n_3, n_4) and occurs with probability p_{n_2,n_3,n_4}, its ordinary degree is recovered from $k = n_2 + 2n_3 + 3n_4$; which is a Diophantine condition [59]. In this introductory setting, we assume that the permissible motifs are cliques of various sizes $\{2, 3, \ldots, m\}$; however, it is important to note that the GCM can be applied to any subgraph topology.

To construct a GCM network, each of the vertices is assigned a tuple of integers, distributed according to the joint degree distribution, that describes their membership in each clique. A set is created for each motif topology, and each vertex is inserted once for each independent motif it is a member of. For instance, if a vertex belongs to 2 2-cliques, 1 3-clique and 0 4-cliques, it is inserted twice into the 2-clique set, once into the 3-clique set and zero times into the 4-clique set. Each set is then used sequentially to construct the network. For a given topology, members are randomly drawn from the set and connected together in the required manner. Once all motifs have been created the process terminates and a particular realisation of the joint degree sequence is obtained, see Fig 1.4. For example, when constructing a 3-clique, three members are drawn at random from the set of vertices involved in triangles, and are connected together in the correct manner.

In GCM networks, cycles are still independent of one another (edge-disjoint), in much the same way that simple edges are in the original model. This means that the accidental formation of a 4-clique during triangle construction through the choosing of two vertices that are already involved in a triangle vanishes with large and sparse networks. Thus, each motif in a GCM network regenerates the locally tree-like property of the ordinary configuration model when considering connections to vertices outside the pre-defined subgraphs. The probability of edge-sharing between independent motifs is dependent on the number of vertices and triangles in the subgraphs for a given number of motifs, however, so this result is valid only for $V \to \infty$.

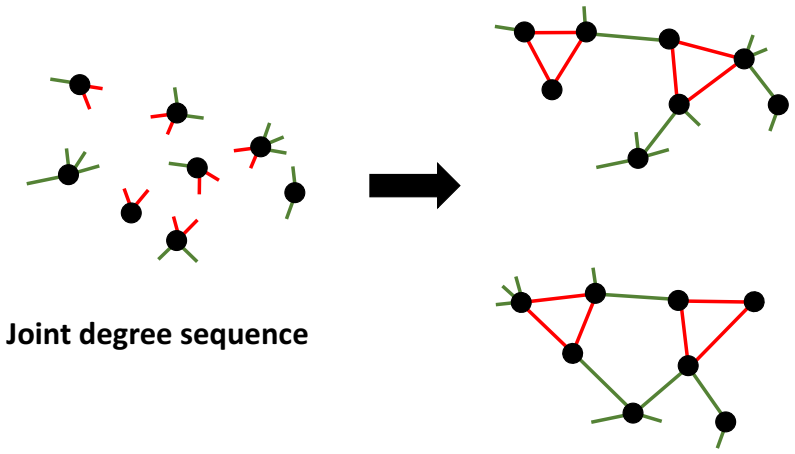


Figure 1.4: In the generalised configuration model a joint degree sequence (left) of stub edges (green) and stub triangles (red) is created by assigning joint degrees, drawn from a joint degree distribution, to a set of vertices. The stubs of each topology are connected together at random (right). The connective structure of each realisation can vary following this procedure. Generating functions (section 2) describe the properties of the entire ensemble of networks for a given joint distribution. Figure inspired by [20]

The degree sequence of a configuration model network is a sequence of tuples

$$(n_{2,1}, n_{3,1}, \ldots, n_{m,1}), \ldots, (n_{2,V}, n_{3,V}, \ldots, n_{m,V}), \tag{1.4}$$

where $n_{\eta,i}$ is the number of motifs of topology η that vertex i is a member of for some set of clique motifs $\{2,3,\ldots,m\}$. As with ordinary edges, not all sequences lead to the successful creation of networks and we now consider necessary conditions on a joint degree sequence in order that it is graphic. The following analysis is taken from Mann *et al* [36]. It is natural to separate the degree tuples and order the joint sequence along each topology as $n_{2,1} \geq n_{2,2} \geq \cdots \geq n_{2,N}$ for the ordinary edges, $n_{3,1} \geq n_{3,2} \geq \cdots \geq n_{3,V}$ for the triangles (and so on) such that

$$\begin{aligned} n_{2,1} &\geq n_{2,2} \geq \cdots \geq n_{2,V} \\ n_{3,1} &\geq n_{3,2} \geq \cdots \geq n_{3,V} \\ &\vdots \\ n_{m,1} &\geq n_{m,2} \geq \cdots \geq n_{m,V} \end{aligned} \tag{1.5}$$

$$\tag{1.6}$$

It is clear that the EGT (the EGI and the HL) must still hold among the overall degrees of the model for the joint degree sequence to be graphic. However, the EGT is no longer sufficient to ensure the graphicality of joint degree sequences according to the GCM. For example, consider the following ordered joint degree sequence for 3 vertices describing ordinary edge and triangle membership, $\{(n_2,n_3)\} = \{(0,1),(1,0),(1,0)\}$. This sequence is graphic according to the EGI, (Eq 1.3, and the HL applied to the overall edges), but is not according to the GCM. In the GCM, we require the EGT to hold among the ordinary

edges such that $\sum_i n_{2,i} = 2H$ where $H \in \mathbb{N}$ is the number of ordinary edges and that

$$\sum_{i=1}^{l} n_{2,i} \leq l(l-1) + \sum_{i=l+1}^{V} \min(n_{2,i}, l) \tag{1.7}$$

holds for $1 \leq l \leq V$. For the triangle degree sequence to be graphical, we require that the sum of the number of triangle edges is divisible by 3

$$2\sum_{i=1}^{V} n_{3,i} = 3T \tag{1.8}$$

which is a modified handshaking lemma, as well as a modified inequality

$$2\sum_{i=1}^{l} n_{3,i} \leq l(l-1) + \sum_{i=l+1}^{V} \min(2n_{3,i}, l) \tag{1.9}$$

must hold for $1 \leq l \leq V$. The factor of 2 in Eq 1.9 is due to the each vertex consuming two edges per triangle. Together these conditions extend the Erdős-Gallai theorem to the tree-triangle model, ensuring that the joint degree sequence is graphic. This can now be readily extended to other GCM networks. The necessary conditions for the graphicality of joint degree sequences of configuration models comprising cliques can now be written by exploiting the characteristic size of each clique. Whilst is is easy to convince ourselves that these conditions are necessary conditions for graphicality, we do not, however, know if these are sufficient conditions.

1.7 Computational considerations

In this section we will discuss aspects of the Monte Carlo simulation that has been used to create the networks and run percolation experiments as well as other computational work.

1.7.1 Percolation

Random graph models were created using Networkx, a Python library for complex networks. Clustered networks were created according to the configuration model algorithm discussed previously. In order to find close agreement between the analytical approach of generating functions and simulation, it is necessary to create large networks, of the order of $1e5$ vertices, with sparse connections so that the locally tree-like property holds and we do not create short range loops by accident. To run bond percolation over these networks, one simply iterates the edges of the graph and tests if they are occupied or not, with a fixed probability T, against a random number; removing those edges from the graph if they are unoccupied. Networkx has methods to find the GCC, as well as other useful properties of networks within its API. The complexity of this bond percolation procedure is $O(E^2)$ where E is the number of edges in the network.

This style of percolation experiment is not the most optimal in terms of computing time. In practice, the Newman-Ziff algorithm (NZA) offers superior experimental run times to the conventional percolation algorithm, with an estimated complexity of $O(E)$. In the NZA, a single run conducts simulations for all values of T at once; in contrast to the conventional method. As T is incrementally increased over the unit interval, connected

components are joined together by the addition of bonds to the network. Rather than throwing the occupied clusters away at each value of T, bonds are simply added to a given realisation and the clusters are updated due to the possible merging of existing clusters. The components are represented by trees, where the root of each tree labels it. When two clusters are merged, the root of one is connected to a vertex belonging to the other tree, such that only one root now remains. The size of the cluster is stored at the root calculating the largest component is simply a lookup exercise. This is an example of a union-find algorithm.

Extending the NZA to the case of multiple interacting epidemics, like those discussed in this thesis, has not been performed (yet); in those cases, the standard percolation algorithm was applied to the relevant sections of networks, such as the RG (for cross-immune diseases) or the GCC (for coinfecting diseases), see chapter 7.

1.7.2 Motif finding algorithms

In a separate avenue of research, the problem of taking an empirical network and covering it with edge-disjoint cliques arose, see chapter 6. For this purpose, we derived a novel clique covering algorithm, which we called the *motif preserving clique cover* MPCC; as well as implementing other covers from the literature [6].

1.7.3 Fixed-point equations

The generating function method yields either a single or, more generally, a system of self-consistent non-linear coupled equations of the form $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$x = f(\boldsymbol{x}) \tag{1.10}$$

where $x \in \boldsymbol{x}$ is an element of vector $\boldsymbol{x}$. Such a system can be converted into a homogeneous form, $x - f(\boldsymbol{x}) = 0$ and solved using a non-linear optimiser. We remark that it is often that case that simple fixed-point iteration is sufficient to find a solution for suitably initialised x; we found $x = 0.5$ to work well since the solution is unique by Jensen's inequality.

1.7.4 Derivatives

Often, we are required to evaluate the coefficients a_k of an infinite series, $f(x) = \sum_k a_k x^k$. To achieve this, we can find derivatives of the infinite series and evaluate them at $x = 0$ to recover a_k as

$$a_k = \frac{1}{k!} \frac{d^k}{dx^k} f(x) \bigg|_{x=0} \tag{1.11}$$

If an analytical solution is not available in closed-form, the derivatives can be performed numerically or symbolically; however, small errors can compound and lead to incorrect results. Newman and Moore [43] indicate that performing a numerical contour integral over the unit circle via the Cauchy formula is the appropriate method

$$\frac{1}{k!} \frac{d^k}{dx^k} f(x) \bigg|_{x=0} = \frac{1}{2\pi i} \oint \frac{f(\xi)}{\xi^{k+1}} d\xi \tag{1.12}$$

1.8 Chapter summary

In this chapter we stated the purpose of this thesis: to study the properties of clustered networks that undergo dynamical processes. We introduced complexity as a topic that transcends traditional disciplines, both in application and method. We discussed random graph ensembles and distinguished between trees, tree-like networks and clustered networks by the inclusion of short range loops that induce correlation among the vertices. We outlined the configuration model as a means to create a particular random network with a prescribed degree distribution. This was then extended to the generalised configuration model that affords the creation of clustered networks. Finally, we introduced the bond percolation process and discussed its criticality, as well as equating it to the final state of the SIR epidemic model.

CHAPTER TWO

GENERATING FUNCTIONS

In this section will discuss the use of generating functions to obtain the properties of a network following bond percolation. The use of generating functions to find solutions to the bond percolation problem originated (we believe) with Fisher and Essam in 1961 [13] where they derived an exact solution for site and bond percolation on Cayley trees. Within the network science community, the next breakthrough was found by Callaway, Newman, Strogatz and Watts in 2000 [7] which provided a solution for general degree distributions. A series of key papers around this time had an enormous influence on the style and direction of research in network science over the next decade [55, 46].

Most of this chapter is background material with the exception of section 2.3, which, although has been derived independently during the PhD, is unlikely to be the first time these functions have been considered. Similarly, the extension of the clustering coefficient to GCM clique networks was developed independently but is likely to have been considered previously.

2.1 Generating the degree distribution

A generating function is a mathematical object that collects the terms of an infinite sequence into the coefficients of a formal power series. The ordinary generating function of a sequence a_n is given by

$$G(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \dots \tag{2.1}$$

Generating functions provide a systematic way to count combinatorial objects. The coefficients of the generating function can be recovered by differentiation

$$a_k = \frac{1}{k!}\frac{d^k}{dx^k} G(x) \tag{2.2}$$

Generating functions, and their manipulation, are commonplace analytical tools in network science to obtain structural properties of graphs. Their utility in this area arises since many of the problems faced in network science are enumerative; often the counting of all possible combinations of a particular configuration or state is required. For instance, suppose that the probability distribution for the fraction of vertices in a network with non-negative integer degree k is p_k, then, the generating function for the distribution of degrees is

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k \tag{2.3}$$

Since the coefficients of this generating function constitute a normalised probability distribution, evaluating this function at $x = 1$ we have $G_0(1) = 1$. Further, since it is positive definite, $G_0(x)$ is absolutely convergent for all $|x| \leq 1$.

Let us examine $G_0(x)$ in more detail. Consider a network of N vertices which each have one of three degrees: k_1, k_2 or k_3 with probabilities p_{k_1}, p_{k_2} and p_{k_3}, respectively. Let us imagine that there has been an epidemic among the population, and that the probability that an edge transmits infection is g, which for now, we assume is a known quantity. Let us pick a vertex at random from the network and examine the probability that it is an infected. Assuming that the network is locally tree-like, the infection probabilities of a vertex's edges are independent of one another; there are no correlations. The probability if choosing a vertex of degree k_1 is p_{k_1}, and the probability that all k_1 of its contact fail to infect is simply the power of single-edge probabilities, g^{k_1}; since, the edges are independent of one another. Therefore, the probability that we choose an uninfected degree k_1 vertex is

$$p_{k_1} g^{k_1} \tag{2.4}$$

Similarly, if we had chosen a vertex of degree k_2 or k_3, we can write the probability that they also fail to be infected. The total probability that the vertex we had chosen from the network at random failed to be infected is then the sum that all vertices we could choose that were uninfected, which is

$$G_0(g) = p_{k_1} g^{k_1} + p_{k_2} g^{k_2} + p_{k_3} g^{k_3} \tag{2.5}$$

Clearly, this is a very useful quantity. The generating function has allowed us to collect the information on the topology of the network by knowledge of the degree distribution and relate it to dynamical quantity g, in turn allowing us to calculate a global property that is

a function of network structure and the dynamics of the process. This is a fundamental method in generating functions: enumerate the properties of a single vertex before scaling it to the global properties of the network.

As another example of the utility of the generating function method for the calculation of network properties, consider the first derivative of $G_0(x)$

$$\frac{d}{dx}G_0(x) = \sum_{k=0}^{\infty} k p_k x^{k-1} \tag{2.6}$$

When evaluated at $x = 1$, this is simply the average degree of the network $G_0'(1) = \langle k \rangle$. Higher moments can also be obtained and in general we have

$$\langle k^m \rangle = \sum_k k^n p_k \tag{2.7}$$

which can readily be found by applying the operator $\left(x\frac{d}{dx}\right)^m$ to $G_0(x)$ evaluated at unity.

Many other properties of a network, such as: the distribution of component sizes, the average number of neighbours at distance l, graph diameter, average path length, clustering coefficient, correlation structure and so on, can be extracted by conducting mathematical operations on generating functions.

Ordinary generating functions can be generalised to coefficients with multiple indices; for instance, a bivariate generating function is written as

$$G(x,y) = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} a_{kl} x^k y^l \tag{2.8}$$

Consider a network whose vertices are arranged into both ordinary edges and edge-disjoint triangles. The joint probability distribution for the fraction of vertices in a network with n_2 ordinary edges and n_3 triangles is p_{n_2,n_3}. The generating function for this distribution is

$$G_0(x,y) = \sum_{n_2=0}^{\infty}\sum_{n_3=0}^{\infty} p_{n_2,n_3} x^{n_2} y^{n_3} \tag{2.9}$$

The average number of ordinary edges the average vertex is connected to is given by

$$\frac{d}{dx}G_0(x,y)\bigg|_{x=1,y=1} = \sum_{n_2=0}^{\infty}\sum_{n_3=0}^{\infty} n_2 p_{n_2,n_3} x^{n_2-1} y^{n_3}\bigg|_{x=1,y=1} \tag{2.10}$$

whilst the average number of triangles the average vertex belongs to is

$$\frac{d}{dy}G_0(x,y)\bigg|_{x=1,y=1} = \sum_{n_2=0}^{\infty}\sum_{n_3=0}^{\infty} n_3 p_{n_2,n_3} x^{n_2} y^{n_3-1}\bigg|_{x=1,y=1} \tag{2.11}$$

The average overall degree, the number of edges that a vertex has when clique membership is ignored, is then

$$\langle k \rangle = \langle n_2 \rangle + 2\langle n_3 \rangle \tag{2.12}$$

Similarly, the generating function for the joint probability distribution, $p_{n_2,n_3\cdots,n_m}$ of a

network whose vertices belong to edge-disjoint cliques up to some size m is

$$G_0(x_2, x_3, \ldots, x_m) = \sum_{n_2=0}^{\infty} \sum_{n_3=0}^{\infty} \cdots \sum_{n_m=0}^{\infty} p_{n_2,n_3\cdots,n_m} x_2^{n_2} x_3^{n_3} \cdots x_m^{n_m} \tag{2.13}$$

where n_m is an index over the number of m-cliques per vertex. The average overall degree is then

$$\langle k \rangle = \sum_{\eta=2}^{m} (\eta - 1)\langle k_\eta \rangle \tag{2.14}$$

where η is an index over clique sizes $\eta = 2, 3, \ldots, m$, $\eta - 1$ is the number of edges per vertex that belong to an η-clique and $\langle n_\eta \rangle$ is the number of independent η-cliques that a vertex will belong to on average in the network. The network model does not need to be limited to clique subgraphs; any edge-disjoint subgraph can be used as a basis for a set of motif topologies, see section 5. However, as we will see later, cliques have a closed-form expression for their percolation properties required for calculation. This is due to two factors: i) the removal of one vertex and its edges from an m-clique generates a clique of size $(m-1)$; and ii) all possible edges between each vertex pair are present in a clique. These properties are not held by any other type of motif, even k-regular subgraphs. (See section 5 for a discussion of these points.)

2.2 Generating the excess degree distribution

The philosophy of the generating function approach for the elucidation of network properties is to determine the k-th term of a probability generating function, $G_0(x)$, in detail. Once the k-th term is known, $G_0(x)$ can be constructed and the properties of the network can be obtained using the machinery developed in section 2.1.

This analysis is performed by selecting a *focal vertex*, a vertex that is chosen at random from the network and examining the properties of its neighbours. Often, in order to determine some information (for now we leave the nature of this information quite abstract) of a particular vertex, information from a neighbour is required. Under the locally tree-like approximation, the properties of the neighbours of a particular focal vertex are independent of one another; they are independent and identically distributed (iid). Therefore, once information associated to one of the neighbours has been found, all k edges of a degree k focal vertex can also be determined. This iid property of tree-like networks transforms the determination of neighbour properties into the determination of the properties of a vertex that is reached by randomly choosing an edge from the network and following it in a randomly chosen direction.

The probability distribution of ordinary edges of a vertex that is obtained by traversing a randomly selected edge in a random direction from the network is not equivalent, in general, to the ordinary degree distribution. This is seen clearly by the zeroth term of the degree distribution, p_0, there is no manner to select a degree zero vertex by following an edge; since, it has no edges to traverse.

For networks comprised of entirely ordinary edges, the excess degree distribution of a vertex is defined as the distribution of edges that a vertex has minus 1 (discounting the edge that was used traversed to reach it). Its generating function can be calculated from

the following expression using the generating function of the ordinary degree distribution

$$G_1(x) = \left[\frac{d}{dx}G_0(x)\right]^{-1}_{x=1} \frac{d}{dx}G_0(x) \tag{2.15}$$

The expression can be simplified using $G_0'(1) = \langle k \rangle$ to obtain

$$G_1(x) = \frac{1}{\langle k \rangle} G_1'(x) \tag{2.16}$$

The excess degree distribution can then be recovered by differentiating according to Eq 2.2. For networks that are comprised of edge-disjoint clique subgraphs, it follows that the probability distribution of clique membership of vertices reached by random edge traversal is dependent on the type of edge that is traversed, see Fig 2.1.

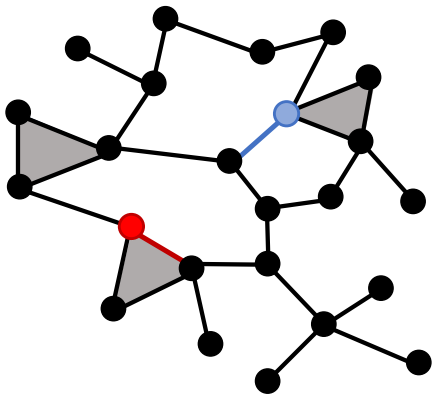


Figure 2.1: A network constructed from 2- and 3-cliques; higher-order structure (such as the 5-cycle) is ignored. If the edge we select at random happened to be the blue 2-clique edge, which was then followed to the blue vertex, we know the vertex *must* belong to at least one 2-clique. If we had selected the red 3-clique edge instead, and followed it to the red vertex, we know the vertex must belong to at least one 3-clique. The distribution of 2- and 3-cliques at a vertex obtained by selecting and traversing a randomly chosen edge depends on the topology of the subgraph it belongs to.

For instance, the distribution of ordinary edges and triangles of a vertex selected by edge traversal in a graph comprised of edge-disjoint ordinary edges and triangles depends on whether the edge that was followed was an ordinary edge or belonged to a triangle. The neighbour must belong to at least one of the cliques that was followed to reach it, whilst there is no restriction on the membership to other cliques. Therefore, there is an excess degree required for each permissible edge-type that is defined in the model. For instance, for a random graph model that comprises edge disjoint 2- and 3-cliques, there are two associated generating functions; one that generates the excess degree distribution for vertices reached from an ordinary edge (the 2-clique) and the other for the excess degree

distribution for vertices reached by traversing a triangle edge

$$G_{1,2} = \left[\frac{d}{dx_2}G_0(x_2,x_3)\right]^{-1}_{x_2=1,x_3=1} \frac{d}{dx_2}G_0(x_2,x_3) \tag{2.17a}$$

$$G_{1,3} = \left[\frac{d}{dx_3}G_0(x_2,x_3)\right]^{-1}_{x_2=1,x_3=1} \frac{d}{dx_3}G_0(x_2,x_3) \tag{2.17b}$$

where the η notation $G_{1,\eta}$ denotes the clique size $\eta = 2,3$.

In the general model with all clique sizes up to size m, there are $m-1$ excess degree distributions $G_{1,\eta} : \mathbb{R}^{m-1} \to \mathbb{R}$ and hence $m-1$ generating functions given by

$$G_{1,\eta} = \left[\frac{d}{dx_\eta}G_0(x_2,\ldots,x_\eta,\ldots,x_m)\right]^{-1}_{x=1} \frac{d}{dx_\eta}G_0(x_2,\ldots,x_\eta,\ldots,x_m) \tag{2.18}$$

with $\eta = 2,\ldots,m$ and where the notation $x = 1$ in the derivative indicates it is evaluated at $(x_2,x_3,\ldots,x_m) = (1,1,\ldots,1)$.

2.3 Specific examples of generating functions

In this short section we consider some specific examples of the generating functions and methods that we frequently use throughout the thesis. Within this section, we consider $\boldsymbol{\tau} = \{2,3,\ldots,m\}$ to be a set of clique sizes that are defined in the model. For instance, a random graph that is composed of 2-, 3- and 4-cliques would have $\boldsymbol{\tau} = \{2,3,4\}$.

2.3.1 Poisson degree distribution

The first example we will consider is the joint Poisson distribution [32]. In this case, the number, n_τ, of independent τ-cliques a vertex is a member of is drawn from a Poisson distribution with average $\langle n_\tau \rangle$. Since each n_τ is an independent variable this is simply a product of independent Poisson distributions

$$p_{n_2,n_3,\cdots,n_m} = \prod_{\tau\in\boldsymbol{\tau}} e^{-\langle n_\tau\rangle} \frac{\langle n_\tau\rangle^{n_\tau}}{n_\tau!} \tag{2.19}$$

where the product extends over each clique topology. This is generated using Eq 2.13

$$G_0(z_2,\ldots,z_m) = \prod_{\tau\in\boldsymbol{\tau}} e^{\langle n_\tau\rangle(z_\tau-1)} \tag{2.20}$$

since

$$e^{z_\tau\langle n_\tau\rangle} = \sum_{n_\tau=0}^{\infty} \frac{(z_\tau\cdot\langle n_\tau\rangle)^{n_\tau}}{n_\tau!} \tag{2.21}$$

It is clear that in this case $G_{1,\tau}(z_2,\ldots,z_m) = G_1(z_2,\ldots,z_m)\ \forall\tau\in\boldsymbol{\tau}$, which can be shown by evaluating Eq 2.18 using Eq 2.20.

2.3.2 Power-law degree distribution with exponential degree cutoff

Next, consider the case where each subgraph is distributed as a power-law distribution with exponential cut-off [32] of the form

$$p_{n_2,n_3,\cdots,n_m} = C\prod_{\tau\in\boldsymbol{\tau}} n_\tau^{-\alpha_\tau} e^{-n_\tau/\kappa_\tau} \tag{2.22}$$

where C, α_τ and κ_τ are constants. The normalisation constant can be found from the condition $G_0(1,\ldots,1)=1$ as

$$C^{-1} = \sum_{n_2=1}^{\infty}\sum_{n_3=1}^{\infty}\cdots\sum_{n_m=1}^{\infty} \frac{e^{-n_2/\kappa_2}}{n^{\alpha_2}}\frac{e^{-n_3/\kappa_3}}{n^{\alpha_3}}\cdots\frac{e^{-n_m/\kappa_m}}{n^{\alpha_m}} \tag{2.23}$$

which is a multipolylogarithm or the form

$$\mathrm{Li}_{s_1,\ldots,s_k}(z_1,\ldots,z_k) = \sum_{n_1>\cdots>n_k>0}\left(\prod_{j=1}^{k} n_j^{-s_j} z_j^{n_j}\right) \tag{2.24}$$

which is convergent on the disc $|z_\tau|<1\ \forall\tau$. The G_0 and $G_{1,\nu}$ generating functions can then be computed as

$$G_0(z_2,\ldots,z_m) = \frac{\mathrm{Li}_{\alpha_2,\ldots,\alpha_m}(z_2e^{-1/\kappa_2},\ldots,z_me^{-1/\kappa_m})}{\mathrm{Li}_{\alpha_2,\ldots,\alpha_m}(e^{-1/\kappa_2},\ldots,e^{-1/\kappa_m})} \tag{2.25}$$

$$G_{1,\nu}(z_2,\ldots,z_m) = \frac{\mathrm{Li}_{\alpha_2,\ldots,\alpha_\nu-1,\ldots,\alpha_m}(z_2e^{-1/\kappa_2},\ldots,z_me^{-1/\kappa_m})}{z_\nu\mathrm{Li}_{\alpha_2,\ldots,\alpha_\nu-1,\ldots,\alpha_m}(e^{-1/\kappa_2},\ldots,e^{-1/\kappa_m})} \tag{2.26}$$

and when $\kappa_\tau\to\infty\ \forall\tau\in\boldsymbol{\tau}$ we have purely power-law networks. In this case we have

$$G_{1,\nu}(z_2,\ldots,z_m) = \frac{\mathrm{Li}_{\alpha_2,\ldots,\alpha_\nu-1,\ldots,\alpha_m}(z_2,\ldots,z_m)}{z_\nu\zeta(\alpha_2,\ldots,\alpha_{\nu-1},\ldots,\alpha_m)} \tag{2.27}$$

where $\zeta(s_1,\ldots,s_k)$ are the multiple Riemann-zeta values

$$\zeta(s_1,\ldots,s_k) = \sum_{n_1>\cdots>n_k>0}\left(\prod_{j=1}^{k} n_j^{-s_j}\right) \tag{2.28}$$

2.3.3 Degree partitioning

In the next example we examine a method of partitioning an (ordinary) degree distribution into a joint distribution of cliques [22, 20, 31]. This method is distinct from the previous examples as it attempts to cover an existing set of degrees with cliques rather than independently distribute cliques. We believe this approach was first introduced by Gleeson and Hackett for 3-cliques [22, 20]; however, the extension to all clique sizes is (we believe) written here for the first time.

Consider an ordered set $\nu,\tau\in\boldsymbol{\tau}=\{m,\ldots,3,2\}$ of clique topologies arranged in descending size for convenience. Let the number of τ-cliques be n_τ and the probability that an edge belongs to a τ clique be x_τ such that $\sum_\tau x_\tau = 1$. The overall degree distribution,

p'_k, is then partitioned as

$$p_{n_2,n_3,\cdots,n_m} = p'_k \sum_{n_2=0} \sum_{n_3=0} \cdots \sum_{n_m=0} \binom{k}{n_2,n_3,\ldots,n_m} x_2^{n_2} x_3^{n_3} \cdots x_m^{n_m} \delta_{k,D} \tag{2.29}$$

where $\delta_{i,j}$ is the Kronecker delta which ensures that the number of edges in the partitioned sequence is equal to the overall degree k; and, where $D = n_2 + 2n_3 + \cdots + (m-1)n_m$ is a linear Diophantine equation. We then have

$$p_{n_2,n_3,\cdots,n_m} = p'_k \sum_{D=k} \binom{k}{n_2,2n_3,\ldots,(m-1)n_m} \prod_{\tau\in\boldsymbol{\tau}} x_\tau^{(\tau-1)n_\tau} \tag{2.30}$$

The multinomial coefficient can be factored as products of binomial coefficients

$$\binom{k}{n_2,2n_3,\ldots,(m-1)n_m} = \prod_\tau \binom{k-\sum_{\nu<\tau}(\nu-1)n_\nu}{(\tau-1)n_\tau} \tag{2.31}$$

and we then have

$$\begin{aligned} p_{n_2,n_3,\cdots,n_m} = p'_k \prod_{\tau\in\boldsymbol{\tau}\setminus\{2\}} \sum_{n_\tau=0}^{\lfloor (k-\sum_{\nu<\tau}(\nu-1)n_\nu)/(\tau-1)\rfloor} \binom{k-\sum_{\nu<\tau}(\nu-1)n_\nu}{(\tau-1)n_\tau} \\ \times x_\tau^{(\tau-1)n_\tau} \left(1 - \sum_{\phi\in\boldsymbol{\tau}\setminus\{2\}} x_\phi\right)^{k-\sum_\phi(\phi-1)n_\phi} \end{aligned} \tag{2.32}$$

where $\lfloor z \rfloor$ is the floor function, and $0 \leq k - \sum_{\nu<\tau}(\nu-1)n_\nu \leq k$ is the number of available edges (those that do not belong to other cliques given the current decomposition of the overall degree into subdegrees).

As an example, consider the edge-disjoint clique decomposition of an overall degree k vertex into 4-clique, 3-clique and 2-clique subgraphs such that $\boldsymbol{\tau} = \{4,3,2\}$. The degree distribution $p_n = p_{n_2,n_3,n_4}$ is given by

$$p_n = p'_k \sum_{n_4=0}^{\lfloor k/3 \rfloor} \binom{k}{3n_4} x_4^{3n_4} \sum_{n_t=0}^{\lfloor (k-3n_4)/2 \rfloor} \binom{k-3n_4}{2n_3} x_3^{2n_3} (1-x_4-x_3)^{k-3n_4-2n_3} \tag{2.33}$$

With this motif partitioning method, we can investigate the importance of *how* a networks clustering is distributed to its percolation properties. For instance, consider a random graph model composed of 2- and 3- cliques. We might wonder if a doubly Poisson degree distribution in each motif

$$p_{n_2,n_3} = \text{Pois}(n_2)\text{Pois}(n_3) \tag{2.34}$$

has different properties to a model where the overall degrees k are Poisson distributed, but the clustering is then created through the partitioning method

$$p_{n_2,n_3} = \text{Pois}(k) \sum_{n_3=0}^{\lfloor k/2 \rfloor} \binom{k}{2n_3} x_3^{2n_3} (1-x_3)^{k-2n_3} \tag{2.35}$$

where a suitable constraint is placed on the sum $n_2 + 2n_3$ for the doubly distributed case such that the first moment of the overall degree distributions are equivalent.

2.3.4 The degree-δ model

While these networks have identical overall degrees, the degree assortativity can also be tuned [22, 20, 31]. To examine the effect of assortativity on the distribution, we use the degree-δ model to allow control of the degree correlations among the subgraphs. For instance, we can fix the clustering to the low degree sites by

$$p_n^{\mathrm{DEL}} = \begin{cases} p_n \delta_{n_s,1} \delta_{n_t,1} \delta_{n_c,0}, & k = 3, \\ p_n \delta_{n_s,1} \delta_{n_t,0} \delta_{n_c,1}, & k = 4, \\ p_n \delta_{k,n_s} \delta_{n_t,0} \delta_{n_c,0}, & \text{otherwise.} \end{cases} \tag{2.36}$$

In other words, vertices in the network are not clustered unless their overall degree is $k = 3$ or $k = 4$, in which case, they are involved in exactly one 3-clique and one independent edge or one 4-clique and one independent edge. This distribution forces the clustering to be positively assorted towards the periphery of the network yet also connects the clique components to the main graph. Of course, we could contain the clustering to the high-degree vertices instead to obtain graphs with different properties.

2.4 Generating the clustering coefficient

One of the fundamental properties of complex networks is the tendency for edges to cluster together. This property is the *transitivity* of the graph. In mathematics, a relation $\circ$ on a set $\{a,b,c\}$ is transitive if the relations $a \circ b$ and $b \circ c$ also imply $a \circ c$. For instance, equality is transitive, $a = b$, $b = c$ and hence $a = c$. In networks, transitivity indicates that the presence of an edge between vertices i and j and j and l also indicates an edge between vertices i and l. Transitive relations in empirical networks, particularly social networks, are defining features of the networks structure and lead to different percolation behaviour.

The *clustering coefficient* of a network is a measure of edge clustering for all vertices. In this section we will show how the generating function formulation can be used to recover the clustering coefficient of a random graph model that is composed of clique subgraphs, as Newman performed for 2- and 3-clique graphs [52]. The global clustering coefficient C of a network with V vertices is defined as

$$C = \frac{3N_\Delta}{N_3} \tag{2.37}$$

where N_Δ is the number of triangles in the network and N_3 is the number of connected triples. Consider a random graph that is composed of clique motifs of sizes $\eta \in \{2,3,\ldots,m\}$. The number of triangles that a vertex that belongs to n_2 2-cliques, n_3 3-cliques and so on is

$$N_{\Delta,n_2,\ldots,n_m} = V p(n_2,\ldots,n_m)(n_2 + \cdots + \mu_m n_m) \tag{2.38}$$

where μ_η is the number of triangles that a vertex belongs to as a member of a η-cycle

$$\mu_\eta = \binom{\eta - 1}{2} \tag{2.39}$$

For instance, $\mu_3 = 1$ while a vertex in 4-clique belongs to 3 triangles. The total number of

triangles in the network is found by summing over the joint degree

$$N_\Delta = \sum_{n_2=0}^{\infty} \cdots \sum_{n_m=0}^{\infty} N_{\Delta,n_2,\ldots,n_m} \tag{2.40}$$

The number of connected triples is given by [52]

$$N_3 = V \sum_k \binom{k}{2} p_k \tag{2.41}$$

$$= V \frac{\langle k^2 \rangle - \langle k \rangle}{2} \tag{2.42}$$

where $k = \sum_\eta (\eta - 1) n_\eta$. For instance, when the model consists of 2- and 3-clique subgraphs, we have $k = n_2 + 2n_3$ and the average number of triangles is simply $\langle n_3 \rangle$ In this case, the clustering coefficient is given by

$$C = \frac{2\langle n_3 \rangle}{\langle k^2 \rangle - \langle k \rangle} \tag{2.43}$$

If the clique distribution is Poisson distributed according to Eq 2.20, then we have $G_0(x,y) = e^{\langle n_2 \rangle (x-1)} e^{\langle n_3 \rangle (y-1)}$ and this expression reduces to

$$C_{\text{Poisson}} = \frac{2\langle n_3 \rangle}{2\langle n_3 \rangle + \langle k \rangle^2} \tag{2.44}$$

2.5 Generating the size of the GCC

Chief among the usefulness of generating functions is their ability to determine the size of the GCC in a network that has undergone bond percolation. Initially, we restrict our attention to networks comprised of ordinary edges (or 2-cliques). We recall from section 1 that the absorbing state of a bond percolation process is binary: vertices either belong to the GCC (if one is present) or they do not; instead they belong to the RG. Applying the philosophy of the generating function approach, in order to determine the macroscopic properties of the network ensemble, the microscopic properties must be considered in detail. In this case, the residence of the k neighbours of a degree k vertex, whether they reside in the GCC or in the RG, is the information that enables the size of the GCC to be calculated.

The calculation proceeds as follows. Let u_2 be the probability that a neighbour at the end of a 2-clique edge *does not* belong to the GCC. Consider an edge that connects the focal vertex to one of its neighbours; let $g_2(u_2, T)$ be the probability that the edge fails to connect the focal vertex to the GCC. We use the iid property of the neighbour states to construct the probability that all of the edges of a degree k vertex fail to connect it to the GCC by raising g_2 to the power of k. The probability that an edge is selected and traversed at random to reach a vertex of excess degree k is generated by $G_1(x)$. Therefore, the probability that the vertex reached by traversing an edge does not belong to the GCC is $G_1(g_2)$. However, by the iid property of neighbour states, this is simply the probability that a neighbour is in the RG of the percolation process. Hence, we can write

$$u_2 = G_1(g_2) \tag{2.45}$$

This is a self-consistent expression that can by solved by fixed point iteration starting from a suitable initial guess. When there is no GCC present in the network, the only root is the trivial solution $u_2 = 1$. This corresponds to the case when the probability that the neighbour is not connected to the GCC is unity. At the critical point, a GCC forms in the network and the solution bifurcates to yield an additional fixed point in the unit interval. Once u_2 has been calculated, the expectation value for the fraction of the network occupied by the GCC, S, is given by

$$S = 1 - G_0(g_2) \tag{2.46}$$

Generalising this to the clique subgraph model, a probability u_η is introduced for each edge-type that could be traversed from a focal vertex to a neighbour that represents the probability that the neighbour that is reached is not connected to the GCC. Due to the presence of the interconnecting clique edges, the residence states of neighbours within a given clique are no longer iid. For instance, if one of the vertices in a clique is attached to the GCC, it is more likely that other vertices within the clique are also connected when compared to degree equivalent vertices with ordinary edges, simply due to the additional edges between the neighbours. The probability that a focal vertex is not connected to the GCC despite its membership in a η-clique subgraph is $g_\eta^{\eta-1} = g_\eta^{\eta-1}(u_\eta, T)$; which is the probability that all of the $\eta - 1$ edges fail to connect it. By containing the neighbour correlations induced by each clique into $g_\eta^{\eta-1}$, the iid property between independent cliques is recovered. Thus we have the following system of coupled equations

$$u_2 = G_{1,2}(g_2, g_3^2, \ldots, g_m^{m-1}) \tag{2.47a}$$
$$u_3 = G_{1,3}(g_2, g_3^2, \ldots, g_m^{m-1}) \tag{2.47b}$$
$$\vdots \tag{2.47c}$$
$$u_m = G_{1,m}(g_2, g_3^2, \ldots, g_m^{m-1}) \tag{2.47d}$$

Similarly to the simple case, the system has a trivial root at the fixed point $u_2, u_3, \ldots, u_m = (1, 1, \ldots, 1)$ that bifurcates in each dimension as the GCC emerges. Performing a linear stability analysis at this fixed point leads to the elucidation of the critical bond occupancy probability.

The expectation value for the size of the GCC in the clique model is then given by

$$S = 1 - G_0(g_2, g_3^2, \ldots, g_m^{m-1}) \tag{2.48}$$

2.6 Equivalent expressions of g_2

We now discuss the details of how g_2 can be calculated. The probability, g_2, that the edge fails to connect the focal vertex to the GCC is given by the sum of all the probabilities that allow it to fail to do so. However, as we see below, there is more than one way to enumerate g_2 and the interpretation of the enumeration procedure is a point of philosophical debate which we would like to explore. The expressions in this section are all due to Newman from various papers that span over a decade and whilst they are numerically equivalent, the structure, the motivating logic as well as implicit assumptions regarding the residence state of the focal vertex are not equivalent in all cases. We might ask is there a representation with a particular benefit; or equivalently, does a particular interpretation yield insight that the others do not?

The exact closed-form expression for g_2 was introduced by Newman [45] and is given by

$$g_2 = 1 - T + u_2 T \tag{2.49}$$

This formula includes the probability that the edge was unoccupied by the percolation process, $1-T$; and, the probability that the edge *was* occupied, but the neighbour did not belong to the GCC, $u_2 T$. It can be concluded from this expression that the focal vertex must itself be unattached to the GCC; since, the neighbour can not be unattached and the edge to the focal vertex occupied simultaneously. Thus, the expression enumerates the possible ways for a focal vertex to reside in the RG.

Consider again that the focal vertex was unattached to the GCC. The neighbouring vertex could be unattached, u_2, or it might be attached but then failed to occupy the edge, $(1-u_2)(1-T)$, such that the focal vertex remains unattached to the GCC with probability

$$g_2 = u_2 + (1-u_2)(1-T) \tag{2.50}$$

We infer that this expression *apriori* assumes that the focal vertex belongs to the RG because the occupation state of the edge that connects to the unconnected neighbour (the first term, u_2). For instance, the edge could be occupied, T, or unoccupied, $1-T$, which is equivalent to multiplying the term by unity, $u_2(1-T+T)$. This expression was used by Newman to model two seasonal epidemics where the second disease spreads on the RG created by the first [49]. In this case only the uninfected neighbours can be infected by the second disease. This expression was used because it isolates this term directly.

Next, consider that the focal vertex itself is in the GCC. There are now three neighbour types that surround this embedded vertex: unattached neighbours, attached neighbours (by vertices other than the focal vertex) and attached neighbours (that the focal vertex directly attached). These three possibilities sum respectively to yield the following expression for g_2

$$g_2 = u_2(1-T) + (1-u_2)(1-T) + u_2 T \tag{2.51}$$

This expression was considered by Newman and Ferrario [53] to model two seasonal epidemics where the second disease spreads on the GCC created by the first. In this case, the infected neighbours can be sources of the second disease. This expression isolates these with particular focus on the infection pathways in the GCC.

Lastly, we consider another expression that was first introduced by Newman and Ferrario [53]. The probability that an edge fails to connect the focal vertex to the GCC is given by 1 minus the probability that the edge successfully connects it to the GCC. This, in turn, is given by the probability that the neighbour was itself connected to the GCC and that the edge was occupied $(1-u_2)T$ such that

$$g_2 = 1 - (1-u_2)T \tag{2.52}$$

This expression does not *apriori* indicate a residence state for the focal vertex; it can belong to the GCC or the RG. It only indicates that the particular edge in question failed to connect the focal vertex.

For a comparison of formulae for g_3 in a similar manner to the analysis conducted for g_2, please see appendix A.

2.7 Generating the critical point

From section 1.3, we understand that the expected size of the GCC undergoes a phase transition from zero to non-zero as the bond occupancy probability is increased in the unit interval. The specific value at which this occurs is the critical threshold T_c. In this section we show how generating functions can be used to find T_c for tree-like networks and examine the critical point of two random graph models.

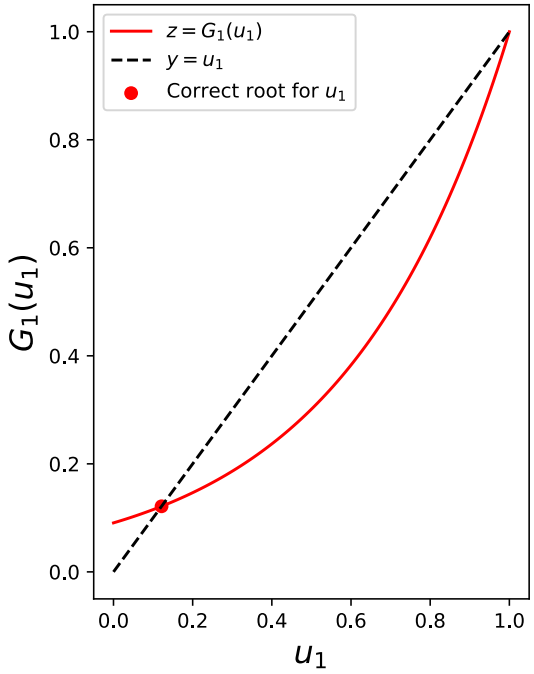


Figure 2.2: A depiction of the graphical solution with the non-trivial root marked with a scatter point where the generating function crosses the $y = u$ dashed line, other than at $u = 1$.

At $T < T_c$ the probability that a neighbour does not belong to the GCC along an ordinary edge, u_2, is exactly unity; since, a GCC does not exist in the network. In other words, $u_2 = 1$ is always a trivial solution of

$$u_2 = \sum_k \frac{k p_k}{\langle k \rangle} g_2^{k-1}(u_2) \tag{2.53}$$

where $g_2(u_2)$ is the probability that an edge fails to connect the vertex to the GCC. This corresponds to the absence of a percolating cluster, $S = 0$, from

$$S = 1 - \sum_k p_k g^k(u_2) \tag{2.54}$$

At the critical point another solution for u_2 appears in the unit interval. This can be determined graphically by plotting $u_2 = F(u_2)$ where $F(u_2) = \sum_k k p_k g_2^{k-1} / \langle k \rangle$ against $y = u_2$. When $u_2 = 0$, we have $F(0) = p_1 / \langle k \rangle$; whilst, at $u_2 = 1$ we have $F(1) = 1$. The first and second derivatives are always positive for $0 < u_2 < 1$ which, by Jensen's

inequality [24], implies that F is monotonically increasing and convex with u_2. The point of intersection between the curve $y = F(u_2)$ and the line $y = u_2$ exists only when the derivative $F'(u_2)$ at $u_2 = 1$ is larger than the derivative of $y = u_2$. This condition is written as

$$\frac{d}{du_2} G_1(g_2)\bigg|_{u_2=1} > 1 \tag{2.55}$$

or, in full we have

$$\frac{d}{du_2}\left(\sum_k \frac{kp_k}{\langle k\rangle} g_2^{k-1}(u_2)\right)\bigg|_{u_2=1} > 1 \tag{2.56}$$

At the point when the derivative *equals* unity, the GCC first forms in the network and the solution for u_2 bifurcates with the presence of a non-trivial root. Performing the derivative, with $g_2(1) = 1$, we arrive at the condition [8] for the presence of a finite sized GCC in the network as

$$T_c = \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle} \tag{2.57}$$

Therefore, the critical bond probability is a function of the topology of the network itself through the first and second moment of the degree distribution. When $T = 1$ we have

$$\frac{\langle k^2\rangle}{\langle k\rangle} > 2 \tag{2.58}$$

This condition is known as the Molloy-Reed criterion [42] and is exact for infinitely sized, uncorrelated tree-like networks. For Poisson networks, using $\langle k^2\rangle = \langle k\rangle(\langle k\rangle + 1)$ we have a giant component if $\langle k\rangle > 1$. Therefore, the average degree has to be larger than 1 for a GCC to exist. For scale-free networks with V vertices, the m-th moment of the power-law degree distribution with exponent α is

$$\langle k^m\rangle = (\alpha-1)k_{\min}^{\alpha-1}\int_{k_{\min}}^{k_{\max}} k^{m-\alpha}dk \tag{2.59}$$

$$= \frac{\alpha-1}{m-\alpha+1}k_{\min}^{\alpha-1}[k_{\max}^{m-\alpha+1} - k_{\min}^{m-\alpha+1}] \tag{2.60}$$

where $k_{\min}$ and $k_{\max}$ are the minimal and maximal degrees, respectively. For $2 < \alpha < 3$ we have

$$\frac{\langle k^2\rangle}{\langle k\rangle} = \frac{\alpha-2}{3-\alpha}k_{\min}^{\alpha-2}k_{\max}^{3-\alpha} \tag{2.61}$$

In the infinite size limit, $V \to \infty$, the degree of the largest hub becomes infinite, $k_{\max} \to \infty$ and hence the second moment diverges for finite $\langle k\rangle$. In this case, the Molloy-Reed criterion is always satisfied and a GCC is always present in the network. For bond percolation, the critical point occurs at $T_c = 0$ from Eq 2.57. However, there could be a non-zero threshold when certain degree correlations are present [3, 4, 44], or if the network is low-dimensional [60], or if transitivity has significant measure [12] .

For random graphs that do contain short range cycles, including networks with clique subgraphs, the standard Molloy-Reed condition fails to predict the critical point. We will examine the critical point of these networks in chapter 4.

2.8 Chapter summary

In this chapter we have reviewed the use of generating functions to describe the structural properties of locally tree-like complex networks as well as properties relating to bond percolation. It was shown that the degree distribution is the fundamental descriptive object required for the formulation along with a microscopic view of the local environment of a randomly selected vertex and its immediate neighbours. We then discussed the generalisation of this model for the case of clustered networks that are constructed according to the GCM algorithm. Due to their simplicity, it was assumed that the subgraphs permitted in the model were cliques of a given size; however, we will broaden this later in Chapter 5. We gave two example degree distributions for GCM networks: joint Poisson and power-law with exponential degree cutoff as well as discussing the partitioning of overall degrees into clique covers via the degree-δ model due to Gleeson *et al*. We also gained a deeper understanding of the critical point of the percolation process and how the generating function formulation can be used to find the critical bond occupation probability or transmissibility. We also introduced a clustering coefficient for GCM graphs and commented on the different representations of the quantity g_2 and their various implications on the component of the network that it belongs to.

CHAPTER THREE

CLIQUE RANDOM NETWORKS

In this chapter we will discuss the enumeration $g_{\eta}^{\eta-1}$ for random graphs composed of edge-disjoint clique subgraphs. In section 3.2 we examine a closed-form formula for all cliques, using 3-cliques as a motivating example of the counting procedure involved. In section 3.3 we introduce an alternative expression for $g_{\eta}^{\eta-1}$ based on a semantic reformulation of the problem that builds on the discussion in section 2.6. We will show in section 7.2 that the expression described in section 3.2 is the correct approach for the study of subsequent percolation events on the RG of a bond percolation process; whilst the new architecture from section 3.3 is the correct description of $g_{\eta}^{\eta-1}$ required for the study of coinfecting epidemics on networks.

3.1 The effect of clique clustering on percolation

The role that clustering plays on the percolation properties, such as the location of the critical point and the size of the GCC, have been the subject of much investigation in the literature [39, 28, 52, 20, 22, 18, 47, 11, 5, 63, 62, 32, 19]. This is because there are a few subtleties in the way in which clustered networks can be constructed; and, without explicit control, they can lead to dichotomous conclusions regarding the effects of clustering. The confusion is best summarised by Miller [39]

> *... a number of studies have investigated the impact of clustering on epidemics. Some found that clustering reduces the sizes of epidemics and raises the epidemic threshold. That is, clustering reduces the size of giant components and raises the percolation threshold. However, others have shown that clustering appears to reduce the threshold.* (Miller, 2009)

Miller shows [39] that the discrepancy occurs due to the associated assortativity among clustered vertices. Vertices involved in triangles tend to have higher degrees than vertices with only ordinary edges. During the construction process of the GCM, by connecting triangles together, inevitably vertices with a high proportion of triangles are segregated from those with a low proportion of triangles. If the overall degrees of vertices with lots of triangles are different from the degrees of vertices with few triangles, then this effect will cause correlation of among the degrees. By constructing clustered and unclustered networks with the same degree distribution and nearest-neighbour degree correlations, Miller found that clustered networks have smaller GCCs and higher critical thresholds than unclustered networks.

When high degree vertices are preferentially connected together, (by either ordinary edges or triangles) the critical threshold is always reduced. This is because it is easier for a connected component of occupied edges to form from a well connected substrate during percolation. Likewise, the finite components are likely to be preferentially composed of low degree vertices. This explains the slower growth of the GCC compared to ordinary configuration model networks with equal degree distributions for $T_c < T < 1$; since, as finite components attach to the GCC, only a few vertices are added to the well connected core. This point is independent of clustering. Gleeson *et al* [20] takes this further by examining the correlation structure beyond the nearest neighbours in the *coloured-edge* model. Whilst supporting Miller's findings, Gleeson also showed that long range correlation structure (beyond nearest neighbour correlations) also influence the properties of the percolation process.

For clustered networks, the size of the GCC reduces with increasing clustering when compared to networks with equivalent correlation structures such that

$$S_{\text{clustered}} \leq S_{\text{unclustered}} \tag{3.1}$$

This happens because triangles contain redundant edges whose presence does not increase the size of the GCC. One in three edges in a triangle are redundant in this way; since, its removal does not isolate a vertex. Thus for a given average degree, and hence a given total number of edges, fewer vertices can be connected together in a network of triangles than in a network of ordinary edges [52]. According to Gleeson [20], the increase in the percolation threshold for clustered networks is a simple consequence of this result

> ... since the GCC size in the clustered network is smaller than or at most equal to that in the unclustered network for all T, the transition point where the clustered GCC size becomes nonzero must be larger than the transition point for the unclustered network. (Gleeson, 2010)

Therefore, we conclude that clustering acts to increase the percolation threshold

$$T_c^{\text{clustered}} \geq T_c^{\text{unclustered}} \tag{3.2}$$

To illustrate the role of larger clique sizes on the percolation threshold and the expected size of the GCC, see Fig 3.1.

3.2 An exact closed-form expression for cliques

Until recently, there were two methods to obtain the $g_\eta^{\eta-1}$ polynomials for random graphs composed of edge-disjoint cliques. These are: i) the exhaustive enumeration of all possible combinations of occupied and unoccupied edges resulting in an equation, which is described as an *exponentially slow* procedure with increasing clique size; or, ii) numerical evaluation by recursion, which is fast, but yields no equation. Newman had previously found the polynomials for clique networks via a recursive method [47] to numerically determine $g_\eta^{\eta-1}$ for a η-clique. Newman's method depends on the probability, $P(k \mid \eta)$, that a particular vertex belongs to a connected cluster of k vertices in an η-clique, including itself. This is given by Eq 7 in [47] as

$$P(k \mid \eta) = \binom{\eta-1}{k-1}(1-T)^{k(\eta-k)}P(k \mid k) \tag{3.3}$$

where we have relabeled Newman's $p \to T$ and $q \to 1-T$ to be in-keeping with our notation. These conditional probabilities are evaluated via recursion from an initial condition of $P(1 \mid 1)$ and

$$P(k \mid k) = 1 - \sum_{l=0}^{k-1} P(l \mid k) \tag{3.4}$$

For the purpose of comparison to our closed-form expression which we develop in this chapter, we have the following equality

$$g_\eta^{\eta-1}\big|_{u_\eta=1} = \sum_{k=1}^{\eta} P(k \mid \eta) \tag{3.5}$$

Mann *et al* [36] recently introduced a closed-form expression for $g_\eta^{\eta-1}$ based on enumerating the ways that a focal vertex can remain unattached to the GCC despite its involvement in a η-clique. The expression results in a polynomial in increasing powers of u_η with coefficients equal to the total probability that the focal vertex fails to be connected to the GCC, multiplied by the number of different ways that failure-mode can occur. For the 2-clique, this method yields Eq 2.49 and hence, we refer to this closed-form expression as the *canonical approach*. Each increasing power assumes that an additional neighbour in the η-clique is itself unattached to the GCC. For example, the first term in the polynomial for an η-clique is given by

$$u_\eta^0(1-T)^{\eta-1} \tag{3.6}$$

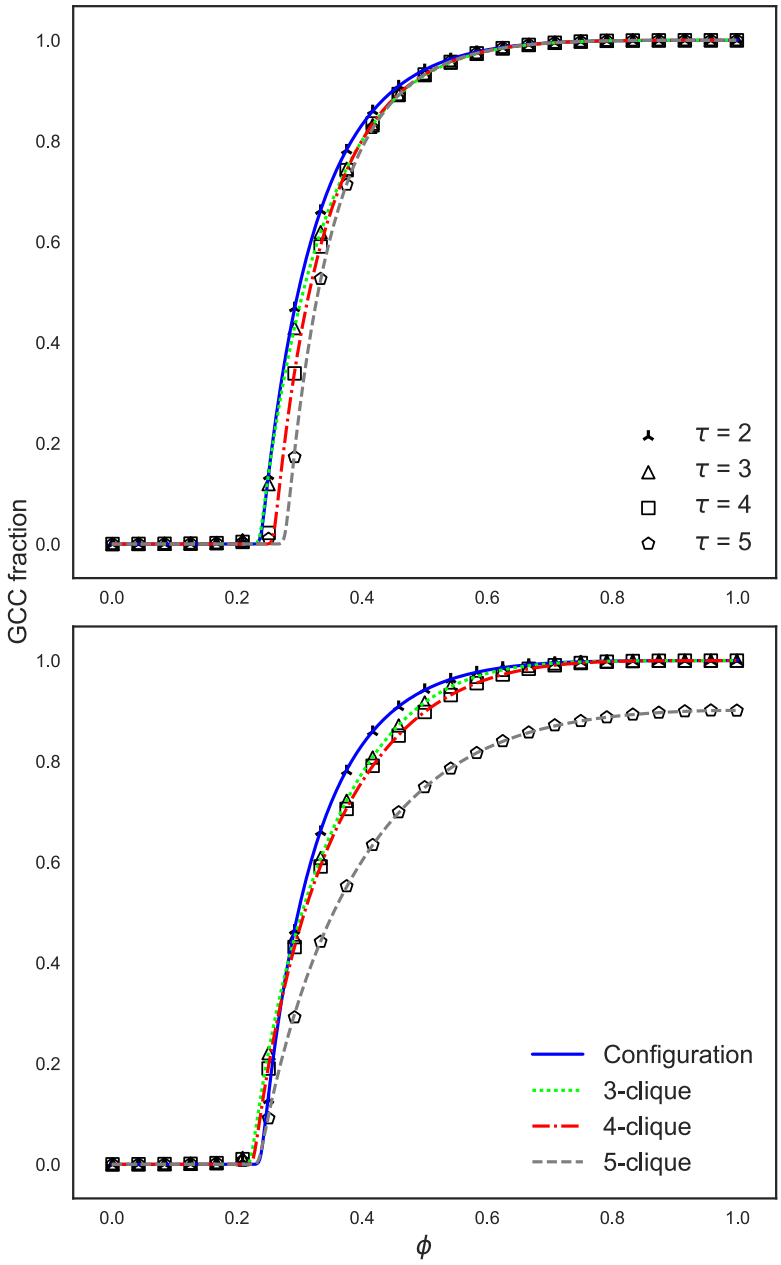


Figure 3.1: The fraction of the network occupied by the GCC as a function of bond occupancy for increasing clique sizes, τ with inverted assortativities. Vertices are either degree 4 or 6 with high (low) degree vertices tending to be clustered in the top (bottom) experiment. From these experiments we observe that higher-order clustering increases the percolation threshold and reduces the size of the GCC. However, the dominant effect depends on the assortativity of the clustering. Scatter points are from Monte Carlo simulation whilst plotted lines are theoretical results. Figure reproduced from [32].

which accounts for the unique mode in which all of the focal vertex's neighbours in the η-clique are attached to the GCC and so all of the edges to the focal vertex must not be occupied in order that it resides in the RG. The next term accounts for the contribution of those cases when there is one neighbour that is also in the RG with the focal vertex and so is linearly dependent on u_η. For this to occur, the edge connecting the focal vertex to this

neighbour is occupied, but all of the remaining edges to the other vertices in the clique are unoccupied. The first occurrence of this mode is for a 2-clique. The linear term occurs with probability uT; since all edges are exhausted, there is no dependence on $(1-T)$. Next, consider this mode for the 3-clique, both of the other edges in the clique are required to be unoccupied, $(1-T)^2$, and there are two ways for this to occur due to the symmetry of the triangle; hence, we obtain $2u_3T(1-T)^2$. For larger cliques, the coefficients of the u_η^1 term depends on the number of ways the focal vertex can connect to a single neighbour and that they both fail to become connected to the GCC. Since there are only 2 neighbour vertices in a 3-clique, the polynomial for a triangle terminates at quadratic dependence on u_3. Together, $g_3^2(u_3,T)$ is given by

$$g_3^2(u_3,T) = (1-T)^2 + 2T(1-T)^2u_3 + (3T^2(1-T)+T^3)u_3^2 \tag{3.7}$$

The clusters for this expression are shown in figure 3.2.

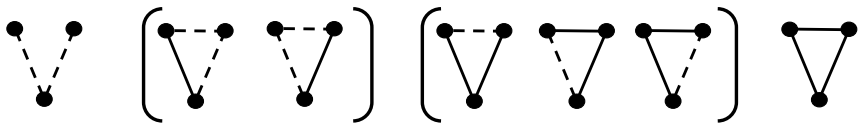


Figure 3.2: A graphical representation of g_3^2 from the closed-form expression in Eq 3.7. Occupied edges are solid lines whilst unoccupied edges are dashed; in each case the focal vertex is the bottom vertex.

The closed-form expression for the evaluation of $g_\eta^{\eta-1}$ now proceeds as follows. Consider a 6-clique whose edges are all occupied and partition the edges into *exterior* (around the outside) and *interior* (through the middle of the clique) sets, see Fig 3.3. For the graph to be connected, and the focal vertex (bottom vertex in Fig 3.3) to be unattached to the GCC, all vertices must themselves be unattached to the GCC; hence the expression for g_6^5 that accounts for these graphs will be proportional to u_6^5. For the fully connected case the probability, $P(0 \mid \eta, 0)$, that the focal vertex remains unattached to the GCC is

$$P(0 \mid \eta, 0) = u^{\eta-1}T^\eta T^{\eta(\eta-1-2)/2} \tag{3.8}$$

where we have partitioned the edges into η exterior edges and $\eta(\eta-1-2)/2$ interior edges. The notation $P(j \mid \eta, r)$ indicates a clique of size η with r vertices attached to the GCC and j edges removed in addition to those that connect to the r removed vertices (see below).

If one of the interior edges is unoccupied, we have

$$P(1 \mid \eta, 0) = q_{\eta,\eta(\eta-1)/2-1}u^{\eta-1}T^\eta T^{\eta(\eta-1-2)/2-1}(1-T) \tag{3.9}$$

where $q_{n,k}$ is the number of connected graphs of n labeled vertices over k edges (see appendix B).

It happens that all of the interior edges and one of the exterior edges can be removed and the graph can remain connected, (Fig 3.3, right). The removal of another edge would isolate a vertex, and so, the term would no longer require all vertices to be unattached; it would be proportional to u_6^4.

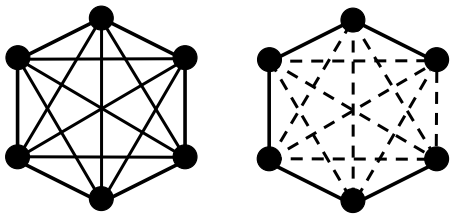


Figure 3.3: All of the interior and one of the exterior edges can be removed from a clique and still make a connected graph among all of the vertices.

The enumeration of the contribution of all graphs that can occur that require u_6^5 can now be written as

$$P(\eta,0) = \sum_{j=0}^{\eta(\eta-1-2)/2+1} P(j \mid \eta,0) \tag{3.10}$$

where

$$P(\eta,0) = \sum_{j=0}^{\eta(\eta-1-2)/2+1} q_{\eta,\eta(\eta-1)/2-j} u^{\eta-1} T^{\eta} T^{\eta(\eta-1-2)/2-j} (1-T)^j \tag{3.11}$$

where $\eta(\eta-1-2)/2+1$ is the number of interior edges plus 1, j is an index of the number of currently removed edges between 0 and all those possible that lead to a connected graph. The notation $P(\eta,0)$ is the probability that an η-clique with 0 removed vertices fails to attach the focal vertex to the GCC.

We now examine the case where a single neighbour vertex within the η-clique is attached to the GCC, see Fig 3.4 (left). There are $(\eta-1)$ vertices that could be attached and all of the edges which connect to this removed vertex must be $(1-T)$, of which there are $(\eta-1)$. This occurs with probability

$$P(0 \mid \eta,1) = (\eta-1) u^{\eta-2} T^{\eta-2} T^{\eta-1)(\eta-1-1-2)/2} (1-T)^{\eta-1} \tag{3.12}$$

Similarly to the previous case, we can remove edges from this graph and still retain connectivity among the $\eta-1$ vertices that belong to the RG. Removal of a single edge occurs with probability

$$P(1,\eta,1) = (\eta-1) q_{\eta-1,X_{\eta-1,1}} u^{\eta-2} T^{\eta-2} T^{(\eta-1)(\eta-1-1-2)/2} (1-T)^{\eta-1+1} \tag{3.13}$$

where $X_{\eta-r,j}$ is the number of edges in the $(\eta-r)$-clique minus j

$$X_{\eta-r,j} = (\eta-r)(\eta-r-1)/2 - j \tag{3.14}$$

The removal of $j \in [0, E(\eta-1)]$ edges, where

$$E(N) = \frac{N(N-1-2)}{2} + 1 \tag{3.15}$$

occurs with probability

$$P(\eta,1)=(\eta-1)\sum_{j=0}^{E(\eta-1)}q_{\eta-1,X_{\eta-1,j}}u^{\eta-2}T^{\eta-2}T^{(\eta-1)(\eta-4)/2-j}(1-T)^{\eta-1+j} \tag{3.16}$$

Removal of any further edges would isolate another vertex and so, we have enumerated all possible graphs that can embed a focal vertex in the RG with $u_\eta^{\eta-2}$ vertices from the η-clique also in the RG. When a second vertex belongs to the GCC in the clique, all

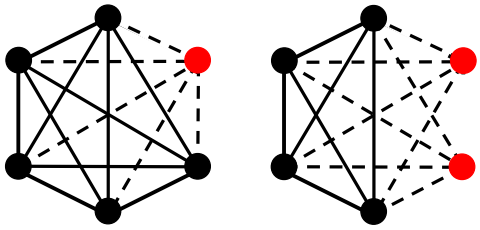


Figure 3.4: (left) A single vertex (red) belongs to the GCC and so there are $\eta-1$ occupied edges to ensure the rest of the subgraph remains in the RG. (right) When two vertices belong to the GCC we do not specify the state of the edges that connect them together.

of the edges that connect this vertex to those in the RG must be $(1-T)$. However, the connections between these removed vertices does need not be unoccupied, see Fig 3.4 (right). Therefore, the total number of edges that are required to be $(1-T)$ is given by the quantity

$$\omega(r)=\sum_{i=1}^{r}(\eta-i)-\frac{r(r-1)}{2} \tag{3.17}$$

where $\sum_{i=1}^{r}(\eta-i)$ is the total number of edges that connect to the removed vertices and $r(r-1)/2$ is the number of vertices that connect removed vertices to each other; the ones that aren't required to be $(1-T)$.

The expression for a clique of size η with two removed vertices is given by

$$P(0\mid\eta,2)=\binom{\eta-1}{2}u^{\eta-3}T^{\eta-3}T^{(\eta-2)(\eta-2-1-2)/2}(1-T)^{2(\eta-2)} \tag{3.18}$$

where $\omega(2)=2(\eta-2)$. All connected graphs among the $\eta-2$ vertices in the RG now occur with total probability

$$\begin{aligned}P(\eta,2)=\binom{\eta-1}{2}&\sum_{j=0}^{E(\eta-2)}q_{\eta-2,X_{\eta-2,j}}u^{\eta-3}T^{\eta-3}T^{(\eta-2)(\eta-2-1-2)/2-j}\\&\times(1-T)^{2(\eta-2)+j}\end{aligned} \tag{3.19}$$

All logic required to extend this expression for all cliques has now been encountered. The total probability that a focal vertex fails to be attached to the GCC in an η-clique is

$$g_\eta^{\eta-1}=\sum_{r=0}^{\eta-1}\sum_{j=0}^{E(\eta-r)}P(j\mid\eta,r) \tag{3.20}$$

which is given by

$$g_{\eta}^{\eta-1} = \sum_{r=0}^{\eta-1} \binom{\eta-1}{r} \sum_{j=0}^{E(\eta-r)} q_{\eta-r,X_{\eta-r,j}} (u_{\eta}T)^{\eta-r-1} T^{E(\eta-r)-1-j} (1-T)^{\omega(r)} \tag{3.21}$$

Further information on the derivation of this expression, its computation and confirmation of its exactness, can be found in [36]. To compare the polynomials from Eq 3.21 to those derived by Newman [47], we set $u = 1$ in Eq 3.21. In section 3.3 we will consider an alternative expression of this quantity.

3.3 An alternative closed-form expression of $g_{\eta}^{\eta-1}$ based on inverse logic

In section 3.2 we reviewed the closed-form expression for $g_{\eta}^{\eta-1}$ introduced by Mann *et al* (2021). In this section, we introduce an alternative exact, closed-form expression for $g_{\eta}^{\eta-1}$ based on the semantic re-interpretation of u_{η}, similar to the discussion in section 2.6.

Consider a 2-clique motif. Let $z_2 = 1 - u_2$ be the probability that the vertex reached by traversing a randomly selected edge *does* belong to the GCC. The probability, f_2, that the edge connects the focal vertex to the GCC is given by $f_2 = z_2 T$. We observe the following relation $g_2 = 1 - f_2$, a manifestation of the property of mutual exclusivity of the binary percolation equilibrium. It happens that z_2 satisfies the following self-consistent expression

$$z_2 = 1 - G_1(1 - z_2 T) \tag{3.22}$$

and the size of the GCC can be calculated as a function of z_2 rather than u_2 as

$$S = 1 - G_0(1 - z_2 T) \tag{3.23}$$

This expression of the percolation problem for random graphs appears in Newman and Ferrario [53] and was discussed in section 2.52. In Fig 3.5, we see the graphical solution for z_2 exhibits a concave appearance, rather than the convex shape of the graphical solution for u_2.

Extending this logic, we next consider the 3-clique. Let $z_3 = 1 - u_3$ be the probability that a triangle the focal vertex is a member of is attached to the GCC. The probability that the focal vertex is also attached to the GCC following bond percolation is

$$f_3^2 = 2T(1-T)^2 z_3 + (3T^2(1-T) + T^3) z_3(2 - z_3) \tag{3.24}$$

where each term is shown graphically in Fig 3.6 and the final bracket is $1 - u_3^2 = z_3(2 - z_3)$. This expression is similar to g_3^2, depicted in Fig 3.2. We notice that the mode that fails to connect the focal vertex to the GCC, $(1-T)^2$, in Eq 3.7 is absent from Eq 3.24, however. The size of the GCC for a mixed 2- and 3-clique network can readily be obtained from these expressions as

$$S = 1 - G_0(1 - f_2, 1 - f_3^2) \tag{3.25}$$

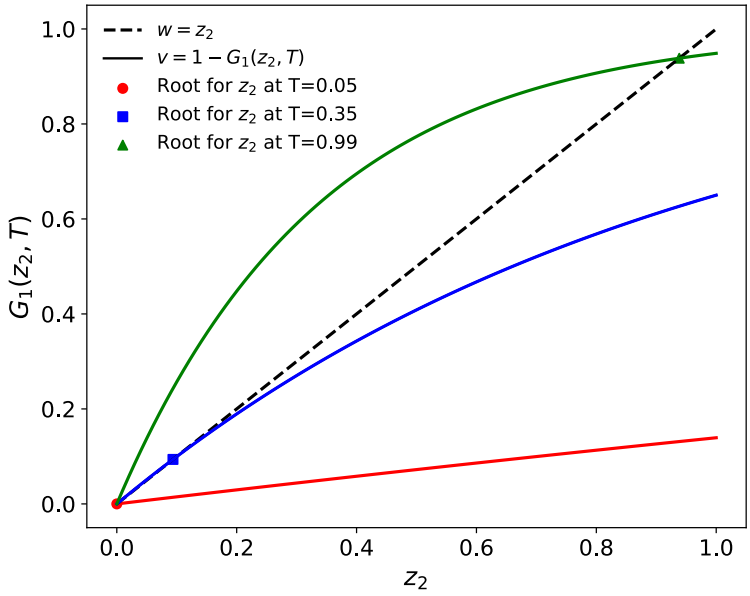


Figure 3.5: The graphical solution of Eq for $z = 1 - u_2$ at three values of T. The value of $z = 0$ is always a fixed point of the system; at the phase transition an additional solution appears. The root for each point is marked with a scatter point and corresponds to the case when $v = w$ and the self-consistent expression for z_2 holds. Unlike the case for u_2, the curves are monotonically concave rather than convex. In this example, an Erdős-Renyi network was used with $\langle k \rangle = 3$.

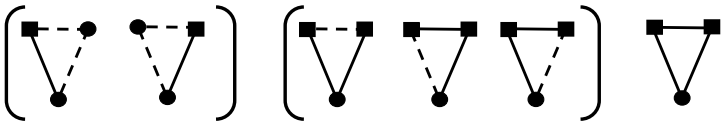


Figure 3.6: A graphical representation of f_3^2 from the closed-form expression in Eq 3.24. Similar to Fig 3.2, occupied edges are solid lines whilst unoccupied edges are dashed; in each case the focal vertex is the bottom vertex and neighbours in the GCC are depicted with squares.

where

$$z_2 = 1 - G_{1,2}(1 - f_2, 1 - f_3^2) \tag{3.26a}$$
$$z_3 = 1 - G_{1,3}(1 - f_2, 1 - f_3^2) \tag{3.26b}$$

The expression for $f_\eta^{\eta-1}$ now follows by simple adaptation of Eq 3.21. Re-labelling $u_\eta \to z_\eta$ and truncating the expression to only remove up to $\eta - 2$ vertices to the RG, such that we always retain at least one connection between the focal vertex and the GCC in the

clique, we have

$$f_{\eta}^{\eta-1} = \sum_{r=0}^{\eta-2} \binom{\eta-1}{r} \sum_{j=0}^{E(\eta-r)} q_{\eta-r,X_{\eta-r,j}} (1-(1-z_{\eta})^{\eta-r-1}) \times T^{\eta-r-1} T^{E(\eta-r)-1-j} (1-T)^{\omega(r)} \tag{3.27}$$

The z_{η} values are computed as

$$z_2 = 1 - G_{1,2}(1-f_2, 1-f_3^2, \ldots, 1-f_m^{m-1}) \tag{3.28a}$$
$$z_3 = 1 - G_{1,3}(1-f_2, 1-f_3^2, \ldots, 1-f_m^{m-1}) \tag{3.28b}$$
$$\vdots$$
$$z_m = 1 - G_{1,m}(1-f_2, 1-f_3^2, \ldots, 1-f_m^{m-1}) \tag{3.28c}$$

Whilst it may seem a tautological change, the concept of inverting the logic from connecting to the RG to connecting to the GCC is an important step for the following section where we define a complement bond percolation problem.

3.4 The complement problem

In section 2.6 we discussed how to calculate the expected size of the GCC for networks composed of 2-cliques. It was found that there is more than one way to derive the quantity g_2, which is the probability that an ordinary edge fails to connect the focal vertex to the GCC. In this section we introduce an additional way to characterise the probability g_2 as well as the complement probability $1-g_2$.

Whilst the expressions for g_2 in section 2.6 all have different motivating logic, each relies on the property of mutual exclusivity of the binary-state percolation equilibrium to find the size of the GCC. For example, the exact closed-form expression calculates the probability that a single edge fails to connect the focal vertex; the probability that a degree k vertex is *not* connected is then simply g_2^k (since the neighbour states are iid over each edge) and hence, the total probability that the average vertex does not belong to to the GCC is $G_0(g_2)$. The probability that the average vertex *does* belong to the GCC is 1 minus this quantity (see Eq 2.46).

In this section, we will consider an alternative method of calculating the size of the GCC by enumerating all connecting pathways to the GCC. This problem is significantly more difficult than the previous methods and constitutes the *complement problem* to the typical method. It is naturally a harder task because there are many ways in which a degree k vertex can be connected to the GCC; whilst there is only one unique way that it fails to be connected. It might be that only $1 \leq k$ of the edges is occupied or it could be that all k edges are occupied. Both scenarios are sufficient to connect the vertex to the GCC and all modes of connection must be accounted for.

Such a scenario is a manifestation of the Anna Karenina principle. The name of the principle derives from Leo Tolstoy's 1877 novel Anna Karenina, whose opening line begins:

> All happy families are alike; each unhappy family is unhappy in its own way.

The same principle was stated much earlier by Aristotle [1]

> Again, it is possible to fail in many ways (for evil belongs to the class of the unlimited, as the Pythagoreans conjectured, and good to that of the limited), while to succeed is possible only in one way (for which reason also one is easy and the other difficult – to miss the mark easy, to hit it difficult); for these reasons also, then, excess and defect are characteristic of vice, and the mean of virtue; For men are good in but one way, but bad in many.

In this case, the success of the generating function method is based on the uniqueness of the manner in which a vertex fails to be connected to the GCC, much like the happy family in Tolstoy's novel.

The complement expression for each degree k vertex will be in the form of a polynomial in u and T. For each k we know that the polynomial must satisfy

$$f_2^k = 1 - g_2^k \tag{3.29}$$

which provides a useful condition to check the exactness of f_2^k once calculated. Consider a vertex of degree $k = 1$ that is connected to the GCC. The probability that the edge is occupied and that the neighbour is connected to the GCC is $(1-u_2)T$. We notice that the consistency condition $1 - g_2 = (1-u_2)T$ holds. For a degree $k = 2$ vertex at least one of the edges must lead to the GCC; however, both edges could connect the focal vertex. We have

$$1 - g_2^2 = [(1-u_2)T]^2 + 2(1-u_2)T(1-T+u_2T) \tag{3.30}$$

which is the sum of the probability that both vertices connect to the GCC and the probability that one edge connects and one edge fails, which can occur in two ways. For $k = 3$ we have

$$\begin{aligned} f_2^k = {} & [(1-u_2)T]^3 + 3[(1-u_2)T]^2(1-T+u_2T) \\ & + 3(1-u_2)T(1-T+u_2T)^2 \end{aligned} \tag{3.31}$$

We leave it as an exercise for the reader to confirm this is equal to $1 - g_2^3$ and we display the first three polynomials graphically in Fig 3.7.

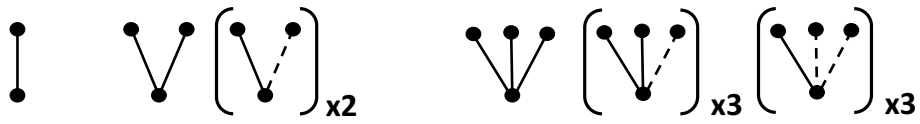


Figure 3.7: A graphical representation of the first 3 polynomials that arise during the inverse calculation procedure for networks composed of ordinary edges. Occupied edges are solid lines whilst unoccupied edges are dashed; in each case the focal vertex is the bottom vertex.

The recipe for a general formula for a degree k vertex is simple: all combinations of connected and unconnected neighbours, subject to the condition that *at least one of them* is occupied, must now be enumerated and multiplied by the number of ways they can occur. We find the general case for a degree k vertex is given by

$$1 - g_2^k = \sum_{l=1}^{k} \binom{k}{l} [(1-u_2)T]^l [1-T+u_2T]^{k-l} \tag{3.32}$$

Notice that the summation index starts at 1, which has the effect of removing the $l = 0$ term from the polynomial.

We can now calculate the u_2 value as

$$u_2 = 1 - \frac{1}{\langle k \rangle} \sum_{k=0}^{\infty} (k+1) p_{k+1} \sum_{l=1}^{k} \binom{k}{l} [(1-u_2)T]^l [1-T+u_2T]^{k-l} \tag{3.33}$$

This expression contains all infection pathways to the GCC, and so its size can be calculated directly rather than using the mutually exclusive trick as the other methods do

$$S = \frac{1}{\langle k \rangle} \sum_{k=0}^{\infty} (k+1) p_{k+1} \sum_{l=1}^{k} \binom{k}{l} [(1-u_2)T]^l [1-T+u_2T]^{k-l} \tag{3.34}$$

We apply this complement counting procedure to cliques in section 3.4.1.

3.4.1 The complement problem for 3-cliques

In section 3.4 we developed an expression for the expected size of a GCC in a random graph composed of ordinary edges without using the mutually exclusive logic that the expressions in section 2.6 rely upon. Here, we extend this approach to enumerate the probability that a focal vertex involved in t edge-disjoint 3-cliques is connected to the GCC. For brevity, we reproduce Eq 3.7 below, which is the probability that a vertex in a triangle fails to be attached to the GCC despite its membership in the triangle

$$g_3^2 = (1-T)^2 + 2T(1-T)^2 u_3 + (3T^2(1-T) + T^3) u_3^2$$

Consider a focal vertex that is connected to a single triangle. The complement polynomial $(f_3^2)^1$ for the probability that the vertex belongs to the GCC is the sum of all combinations of possible connection modes. The expression must be equal to 1 minus the probability that the triangle failed to connect the vertex,

$$(f_3^2)^1 = 1 - g_3^2 \tag{3.35}$$

We find from section 3.3 that the expression is given by

$$(f_3^2)^1 = 2T(1-T)^2(1-u_3) + (3T^2(1-T) + T^3)(1-u_3)^2 \tag{3.36}$$

The reader can assure themselves that this equality is true by inserting the expression for g_3^2 into Eq 3.35. The corresponding polynomial for a vertex that has membership in 2 edge-disjoint triangles is given by

$$\begin{aligned}(f_3^2)^2 =& [2T(1-T)^2(1-u_3) + (3T^2(1-T) + T^3)(1-u_3)^2]^2 \\ &+ 2[2T(1-T)^2(1-u_3) + (3T^2(1-T) + T^3)(1-u_3)^2] \\ &\times [(1-T)^2 + 2T(1-T)^2 u_3 + (3T^2(1-T) + T^3) u_3^2]\end{aligned} \tag{3.37}$$

The general expression for the probability that a focal vertex attaches to the GCC when it

is a member of t triangles is given by

$$1-[g_3^2]^t = \sum_{l=1}^{t}\binom{t}{l}[2T(1-T)^2(1-u_3)+(3T^2(1-T)+T^3)(1-u_3)^2]^l \times[(1-T)^2+2T(1-T)^2u_3+(3T^2(1-T)+T^3)u_3^2]^{t-l} \quad (3.38)$$

We can now calculate the u_3 as

$$u_3 = 1-\frac{1}{\langle t\rangle}\sum_{l=1}^{t}\binom{t}{l}[2T(1-T)^2(1-u_3)+(3T^2(1-T)+T^3)(1-u_3)^2]^l \times[(1-T)^2+2T(1-T)^2u_3+(3T^2(1-T)+T^3)u_3^2]^{t-l} \quad (3.39)$$

The GCC is then given by

$$S = \frac{1}{\langle t\rangle}\sum_{l=1}^{t}\binom{t}{l}[2T(1-T)^2(1-u_3)+(3T^2(1-T)+T^3)(1-u_3)^2]^l \times[(1-T)^2+2T(1-T)^2u_3+(3T^2(1-T)+T^3)u_3^2]^{t-l} \quad (3.40)$$

3.4.2 The complement problem for mixed clique networks

In sections 3.4 and 3.4.1 we developed an approach for calculating the properties of networks that have undergone bond percolation that are composed of 2-cliques or 3-cliques. In this section we continue to investigate the complement problem to account for vertices that can be members of both single edges and triangles. This problem is harder still, as connection to the GCC, if one exists, could occur through any of cliques, regardless of their topology.

For a vertex that belongs to exactly 1 2-clique and 1 3-clique, $(s,t)=(1,1)$, the probability that it is connected to the GCC using the mutually exclusive logic can readily be calculated as

$$P(1,1) = 1-g_2g_3^2 \quad (3.41)$$

where the polynomial $P(s,t)$ indicates membership of s 2-cliques and t 3-cliques. Performing the complement analysis, connection to the GCC could occur via both the 2-clique and the 3-clique, or just one of them, see Fig 3.8. The probability of this is given by

$$P(1,1) = f_2f_3^2+g_2f_3^2+f_2g_3^2 \quad (3.42)$$

where $g_2 = 1-T+u_2T$ and g_3^2 is given by Eq 3.7, both in the canonical form; whilst, f_2 and f_3^2 use the alternative expression derived in section 3.3.

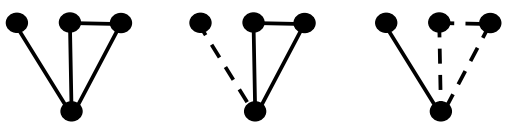


Figure 3.8: The possible modes of connection to the GCC for a focal vertex (bottom) with joint degree $(s,t)=(1,1)$.

The problem can be formalised by introducing a variable X_η for each clique topology, $\eta = 2,3$, that accounts for the number of connections that are made within each topology. For example, for the $(s,t) = (1,1)$ focal vertex in Fig 3.8, the variables associated to each connection mode are $(X_2 = 1, X_3 = 1), (X_2 = 0, X_3 = 1)$ and $(X_2 = 1, X_3 = 0)$. We seek the probability that at least one connection leads to the GCC, which is

$$\begin{aligned} P(\{X_2 \geq 1\} \cup \{X_3 \geq 1\} \mid s,t) &= P(\{X_2 = 0\}^c \cup \{X_3 = 0\}^c \mid s,t) && \text{(3.43a)} \\ &= 1 - P(\{X_2 = 0\} \cap \{X_3 = 0\} \mid s,t) && \text{(3.43b)} \\ &= 1 - P(X_2 = 0 \mid s) P(X_3 = 0 \mid t) && \text{(3.43c)} \end{aligned}$$

where $\{\cdot\}^c$ indicates the complement set. In the final equality, we have recovered the typical complement solution, that uses the mutually exclusive logic from section 2.6 which, written in more familiar notation is simply

$$1 - P(X_2 = 0 \mid s) P(X_3 = 0 \mid t) = 1 - g_2^s g_3^{2t} \tag{3.44}$$

We have

$$1 = \sum_{i=0}^{s} \sum_{j=0}^{t} \binom{s}{i} \binom{t}{j} (f_2)^i (g_2)^{s-i} (f_3^2)^j (g_3^2)^{t-j} \tag{3.45}$$

and therefore,

$$\begin{aligned} P(\{X_2 \geq 1\} \cup \{X_3 \geq 1\} \mid s,t) = \sum_{i=0}^{s} \sum_{j=0}^{t} \binom{s}{i} \binom{t}{j} (f_2)^i (g_2)^{s-i} (f_3^2)^j \\ \times (g_3^2)^{t-j} - g_2^s g_3^{2t} \end{aligned} \tag{3.46}$$

Unlike the expression for ordinary edges in Eq 3.32, we cannot simply start the summations at 1 to remove the zeroth term and so we are forced to manually remove the $i = j = 0$ term explicitly.

The expression for a mixed-clique topology network is then given by the multi-binomial theorem (not to be confused with the multinomial theorem) which deals with products of binomial expressions

$$P(X \mid s, \ldots, n) = \sum_{i=0}^{s} \cdots \sum_{j=0}^{n} \binom{s}{i} f_2^i g_2^{s-i} \cdots \binom{n}{j} f_m^j g_m^{(m-1)(n-j)} - \prod_{l=2}^{m} g_l^{l-1} \tag{3.47}$$

with

$$X = \bigcup_{l=2}^{m} \{X_l \geq 1\} \tag{3.48}$$

and where s is the number of 2-cliques, i is an index over s that counts the number of occupied 2-cliques, n is the number of m-cliques with j indexing occupied m-cliques. The final term accounts for the unique mode where all cliques that the focal vertex belongs to are unoccupied and thus fail to connect it to the GCC.

The calculation of u_2 now proceeds as

$$u_2 = 1 - \frac{1}{\langle s \rangle} \sum_{s=0}^{\infty} \cdots \sum_{n=0}^{\infty} s p_{s,\ldots,n} P(X \mid s-1, \ldots, n) \tag{3.49}$$

and similarly for each $\eta \in [2,m]$. The size of the GCC is given by

$$S = \sum_{s=0}^{\infty} \cdots \sum_{n=0}^{\infty} p_{s,\ldots,n} P(X \mid s,\ldots,n) \tag{3.50}$$

Whilst this may seem a rather complicated way to calculate the GCC, we will see in the coming chapters the benefit of this method for describing interacting epidemic processes on networks using bond percolation.

3.5 Chapter summary

This chapter has concerned the ensemble properties of clique networks that are constructed according to the GCM algorithm. We discussed how subtle differences in the construction process of these networks can lead to networks with globally distinct percolation properties. In order to elucidate those properties, a quantity $g_{\eta}^{\eta-1}$ must be enumerated. Our contribution to the literature was to find a closed-form exact analytical expression for this quantity through a combinatorial enumeration method that counts the connected subgraphs that can be induced on a clique. This expression was formulated around the paradigm of counting the complete set of scenarios in which a randomly selected vertex that belongs to a clique fails to be attached to the GCC. We then showed that this expression can be semantically inverted to yield the probability that the vertex we chose *was* embedded in the GCC. Utilising the dual description, we reformulated how the generating function formulation can be used to describe the percolation properties of clique-clustered networks by introducing the complement problem.

CHAPTER FOUR

COMPONENTS OF CLIQUE RANDOM NETWORKS

In chapter 2.7 we examined the Molloy-Reed condition which marks the critical point of tree-like random networks. When the number of cycles in the network has finite measure, the condition fails to locate the critical point of these networks. In this section we will derive an expression for the critical point of random graphs that comprise of clique subgraphs; extending the Molloy-Reed condition. To achieve this, we extend Newman's work on the critical point of tree and triangle model graphs [52]. Specifically, we provide a derivation of the generating function $h_0(z)$ from first principle combinatorial arguments, before using the expression to find the average component size, which in turn is used to find the critical point of the percolation process. Whilst $h_0(z)$ cannot be evaluated directly, the coefficients can be extracted by Lagrange inversion. We perform this procedure for single topology networks and then extend these results to GCM graphs composed of an arbitrary number of clique topologies. We confirm our formulas by comparing them with Monte Carlo simulations of GCM graphs whose clique membership is Poissonian. We do not draw conclusions on the effect of clustering on the size distribution of the finite components due to the presence of as-yet uncontrolled degree assortativity in the Poisson models; which we hope to investigate further at a future point.

4.1 The distribution of component sizes

Consider a random graph model that is comprised of cliques with sizes $\eta \in 2,3,\ldots,m$ where η is an index over the clique sizes. Let $h_2(z)$ be the generating function for the distribution of number of vertices accessible via the vertex at the end of a single edge; similarly, let $h_3(z)$ be the number of vertices accessible via the vertex at a corner of a triangle and so on for increasing clique sizes. Extrapolating Newman's results for 2- and 3-cliques [52], these distributions are given by self-consistent expressions

$$h_2(z) = zG_{1,2}(h_2(z), h_3^2(z), \ldots, h_m^{m-1}(z)) \tag{4.1a}$$
$$h_3(z) = zG_{1,3}(h_2(z), h_3^2(z), \ldots, h_m^{m-1}(z)) \tag{4.1b}$$
$$= \quad \vdots \tag{4.1c}$$
$$h_m(z) = zG_{1,m}(h_2(z), h_3^2(z), \ldots, h_m^{m-1}(z)) \tag{4.1d}$$

The probability that a vertex chosen at random belongs to a component of a given size is generated by

$$h_0(z) = zG_0(h_2(z), h_3^2(z), \ldots, h_m^{m-1}(z)) \tag{4.2}$$

It is the aim of this section to formally derive this expression from elementary arguments in enumerative combinatorics.

Let the probability that a random vertex has $n_2, n_3, \ldots, n_m$ cliques be given by $p_{n_2,\ldots,n_m}$; and, let $i = 1,\ldots,n_\eta$ be an index over each particular clique of topology $\eta \in [2,m]$. The probability that a particular η clique leads to t accessible vertices is ρ_t. By definition $h_\eta(z)$ is given by

$$h_\eta(z) = \sum_{t=0}^{\infty} \rho_t z^t \tag{4.3}$$

The probability $P(s \mid n_2, n_3, \ldots, n_m)$ that a vertex of joint clique membership $n_2, n_3, \ldots, n_m$ belongs to a component of size s is the probability that the number of vertices reachable along each of its edges sum to $s-1$. We construct this as follows. The probability that the sum of the number of accessible vertices along all edge topologies is s' is given by

$$P(s') = \prod_{\eta=2}^{m} \prod_{i=1}^{(\eta-1)n_\eta} \rho_{t_{\eta_i}} \tag{4.4}$$

The sum of the accessible vertices along each edge of each topology is

$$s' = \sum_{\eta=2}^{m} \sum_{i=1}^{(\eta-1)n_\eta} t_{\eta_i} \tag{4.5}$$

The limits on the summation and product are due to there being $\eta - 1$ edges to be counted per η-clique. To ensure that this sums to the correct value, $s-1$, we use a Kronecker delta

$$P(s) = \delta(s-1, \sum_{\eta=2}^{m} \sum_{i=1}^{(\eta-1)n_\eta} t_{\eta_i}) \prod_{\eta=2}^{m} \prod_{i=1}^{(\eta-1)n_\eta} \rho_{t_{\eta_i}} \tag{4.6}$$

There are many different combinations that make a given component size; we account for

each possible way to achieve a given configuration.

$$P(s \mid n_2,\ldots,n_m) = \sum_{\eta=2}^{m} \sum_{i=0}^{(\eta-1)n_\eta} \sum_{t_{\eta_i}=1}^{\infty} \delta(s-1, \sum_{\eta=2}^{m} \sum_{i=0}^{(\eta-1)n_\eta} t_{\eta_i}) \prod_{\eta=2}^{m} \prod_{i=1}^{(\eta-1)n_\eta} \rho_{t_{\eta_i}} \tag{4.7}$$

For instance, a vertex that belongs to n_2 2-cliques and n_3 3-cliques has

$$P(s \mid n_2,n_3) = \sum_{t_{2_1}=1}^{\infty} \cdots \sum_{t_{2_{n_2}}=1}^{\infty} \cdot \sum_{t_{3_1}=1}^{\infty} \cdots \sum_{t_{3_{n_3}}=1}^{\infty} \delta(s-1, \sum_{i=1}^{n_2} t_{\eta_i} + \sum_{j=1}^{n_3} t_{\eta_j}) \prod_{i=1}^{n_2} \rho_{t_{2_i}} \prod_{j=1}^{n_3} \rho_{t_{3_i}} \tag{4.8}$$

The probability that a randomly chosen vertex belongs to a component of size s and that it has joint clique degree $n_2,\ldots,n_m$ is then

$$P(s,n_2,\ldots,n_m) = p_{n_2,\ldots,n_m} P(s \mid n_2,\ldots,n_m) \tag{4.9}$$

We then average over all combinations of $n_2,\ldots,n_m$ to give the total probability of belonging to a component of size s as

$$\pi_s = \sum_{n_2=0}^{\infty} \cdots \sum_{n_3=0}^{\infty} p_{n_2,\ldots,n_m} P(s \mid n_2,\ldots,n_m) \tag{4.10}$$

Finally, we generate the distribution of component sizes as

$$h_0(z) = \sum_{s=1}^{\infty} \pi_s z^s \tag{4.11}$$

Inserting Eq 4.7 we find

$$h_0(z) = \sum_{\eta=2}^{m} \sum_{n_\eta=0}^{\infty} p_{n_2,\ldots,n_m} \sum_{s=1}^{\infty} z^s \sum_{i=0}^{(\eta-1)n_\eta} \sum_{t_{\eta_i}=1}^{\infty} \delta(s-1, \sum_{\eta=2}^{m} \sum_{i=0}^{(\eta-1)n_\eta} t_{\eta_i}) \prod_{\eta=2}^{m} \prod_{i=1}^{(\eta-1)n_\eta} \rho_{t_{\eta_i}} \tag{4.12}$$

Next, we factor $z^s = z \cdot z^{s-1}$ and evaluate the Kronecker delta

$$h_0(z) = z \sum_{\eta=2}^{m} \sum_{n_\eta=0}^{\infty} p_{n_2,\ldots,n_m} z^{\sum_{\eta=2}^{m} \sum_{i=0}^{(\eta-1)n_\eta} t_{\eta_i}} \sum_{i=0}^{(\eta-1)n_\eta} \sum_{t_{\eta_i}=1}^{\infty} \prod_{\eta=2}^{m} \prod_{i=1}^{(\eta-1)n_\eta} \rho_{t_{\eta_i}} \tag{4.13}$$

In the configuration model, a vertex's edges are independent of one another. This means that if the distribution of a property of a neighbouring vertex is generated by a given generating function, then the distribution of the total property summed over all independent edges is generated by the power of that generating function [54]. More formally, this property is n-ary multiplicative operation on a distribution $f(k)$ known as an n-fold *convolution power*

$$f(k)^{*n} = f(k)^{*(n-1)*f(k)} \tag{4.14}$$

where $f(k)^{*0} = 1$. The convolution power can be expanded into a sum of products of the

form

$$f(k)^{*n} = \sum_{k_1+\cdots+k_n=k} \prod_{i=1}^{n} f(k_i) \tag{4.15}$$

where each $k_i > 0$. If $F(z)$ generates $f(k)$ then $f(k)^{*n}$ is generated by $F(x)^n$. For instance, consider the component distribution of a vertex with two ordinary edges. We can expand the power of the generating function as a double sum of products

$$[h_{1,2}(z)]^2 = \sum_{jk} \rho_j \rho_k z^{j+k} \tag{4.16}$$

$$= \rho_0\rho_0 z^0 + (\rho_0\rho_1 + \rho_1\rho_0)z^1 + (\rho_0\rho_2 + \rho_1\rho_1 + \rho_2\rho_0)z^2 + (\rho_0\rho_3 + \rho_1\rho_2 + \rho_2\rho_1 + \rho_2\rho_0)z^3 + \ldots \tag{4.17}$$

where each power of z is the overall component size whilst the coefficients sum the different ways in which the configuration can be constructed. The convolution power can be applied to multivariate generating functions as

$$f(\boldsymbol{k})^{*n} = f(\boldsymbol{k})^{*(n-1)} * f(\boldsymbol{k}) \tag{4.18}$$

with $f(\boldsymbol{k}) = \delta(\boldsymbol{k})$ and

$$f(\boldsymbol{k}) * g(\boldsymbol{k}) = \sum_{\boldsymbol{j}+\boldsymbol{k}=\boldsymbol{n}} f(\boldsymbol{j}) g(\boldsymbol{k}) \tag{4.19}$$

where the summation is over all partitions of vector n into two parts. We use this fact to write the sum of products in Eq 4.13 as powers of $h_{1,\eta}(z)$ to obtain

$$h_0(z) = z \sum_{n_2=0}^{\infty} \cdots \sum_{n_m=0}^{\infty} p_{n_2,\ldots,n_m} \prod_{\eta=2}^{m} \left[\sum_{t_\eta=1}^{\infty} \rho_{t_\eta} z^{t_\eta} \right]^{(\eta-1)n_\eta} \tag{4.20}$$

$$= z \sum_{n_2=0}^{\infty} \cdots \sum_{n_m=0}^{\infty} p_{n_2,\ldots,n_m} \prod_{\eta=2}^{m} \left[h_{1,\eta}^{\eta-1}(z) \right]^{n_\eta} \tag{4.21}$$

$$= z G_0(h_2(z), h_3^2(z), \ldots, h_m^{m-1}(z)) \tag{4.22}$$

which is the postulated expression. Similar arguments can be made for Eqs 4.1 and we reproduce Newman's expressions when restricted to 2-clique [51] and 2- and 3-clique models [52].

4.2 The mean component size

The expectation value for the average component size in the network is found from the expectation value of $h_0(z)$, which is obtained by taking the derivative with respect to z evaluated at $z = 1$

$$\langle h_0 \rangle = G_0(h_2(1), \ldots, h_m^{m-1}(1)) + z \sum_{\nu=2}^{m} (\nu - 1) \frac{\partial G_0}{\partial h_\nu} \frac{\partial h_\nu}{\partial z} \bigg|_{z=1} \tag{4.23}$$

The derivatives $\partial_z h_\eta(1)$ are given by

$$\left.\frac{\partial h_\nu}{\partial z}\right|_{z=1} = G_{1,\nu}(h_2(1),\ldots,h_m^{m-1}(1)) + z\sum_{\mu=2}^{m}(\nu-1)\frac{\partial G_{1,\nu}}{\partial h_\mu}\left.\frac{\partial h_\mu}{\partial z}\right|_{z=1} \tag{4.24}$$

Inserting the definition of $G_{1,\eta}$ from Eq 2.18 we obtain a Hessian condition

$$\begin{aligned}\left.\frac{\partial h_\nu}{\partial z}\right|_{z=1} &= G_{1,\nu}(h_2(1),\ldots,h_m^{m-1}(1)) \\ &\quad + z\sum_{\mu=2}^{m}(\mu-1)\frac{\partial}{\partial h_\mu}\left[\frac{1}{\langle n_\nu\rangle}\frac{\partial}{\partial h_\nu}G_0(h_2(1),\ldots,h_m^{m-1}(1))\right]\left.\frac{\partial h_\mu}{\partial z}\right|_{z=1}\end{aligned} \tag{4.25}$$

With $h_\mu(1)=1$ from Eq 4.11 we have $G_{1,\nu}(1,\ldots,1)=1$, and so this result can be rewritten as the following matrix equation

$$\boldsymbol{h} = \mathbf{1} + \boldsymbol{\alpha}^{-1}\boldsymbol{H}\boldsymbol{\beta}\cdot\boldsymbol{h} \tag{4.26}$$

where $\mathbf{1}=(1,1,\ldots)$ and $\boldsymbol{h}=(h_2',\ldots,h_m')$ are vectors, $\boldsymbol{H}$ is a Hessian of partial derivatives of $G_0(h_2,\ldots,h_m^{m-1})$ with respect to h_μ and h_ν such that

$$\boldsymbol{H} = \begin{pmatrix} \partial^2_{2,2} & \partial^2_{2,3} & \ldots & \partial^2_{2,m} \\ \partial^2_{3,2} & \partial^2_{3,3} & \ldots & \partial^2_{3,m} \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2_{m,2} & \partial^2_{m,3} & \ldots & \partial^2_{m,m} \end{pmatrix} \tag{4.27}$$

while $\boldsymbol{\alpha}$ is a diagonal matrix of expected values of the number of cliques of a given topology and $\boldsymbol{\beta}$ is a diagonal matrix of the number of direct contacts the focal vertex has per clique such that

$$\boldsymbol{\alpha} = \begin{pmatrix} \langle n_2\rangle & 0 & \ldots & 0 \\ 0 & \langle n_3\rangle & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \langle n_m\rangle \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & m-1 \end{pmatrix} \tag{4.28}$$

Rearranging this equation allows us to solve for the derivatives in Eq 4.23 to find the average component size

$$(\boldsymbol{I} - \boldsymbol{\alpha}^{-1}\boldsymbol{H}\boldsymbol{\beta})\cdot\boldsymbol{H} = \mathbf{1} \tag{4.29}$$

where $\boldsymbol{I}$ is the identity matrix and where diagonal elements are given by $1-\partial^2_{\nu,\nu}/\langle n_\nu\rangle$ and off-diagonal elements are $-\partial^2_{\mu,\nu}/\langle n_\mu\rangle$. For example, consider a random graph with 2- and 3-clique subgraphs. The average component size is given by

$$\langle h_0\rangle = 1 + \langle n_2\rangle h_2'(1) + 2\langle n_3\rangle h_3'(1) \tag{4.30}$$

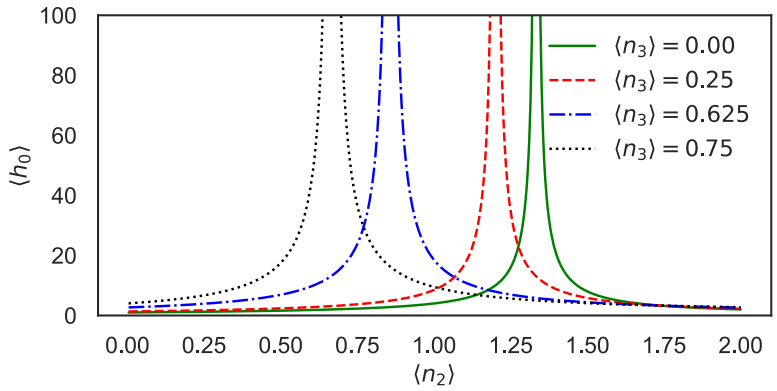


Figure 4.1: The average component size for GCM networks consisting of 2- and 3-cliques whose membership is Poisson distributed with means $\langle n_2 \rangle$ and $\langle n_3 \rangle$, respectively. The average component size diverges at the critical point. This result shows that increasing the number of triangles reduces the critical point; however, this is due to the degree assortativity, rather than the effect of clustering.

The derivatives $h_2'(1)$ and $h_3'(1)$ are given by

$$h_2'(1) = 1 + \frac{1}{\langle n_2 \rangle} \frac{\partial^2 G_0(1,1)}{\partial h_2 \partial h_2} h_2'(1) + \frac{2}{\langle n_2 \rangle} \frac{\partial^2 G_0(1,1)}{\partial h_2 \partial h_3} h_3'(1) \tag{4.31a}$$

$$h_3'(1) = 1 + \frac{1}{\langle n_3 \rangle} \frac{\partial^2 G_0(1,1)}{\partial h_3 \partial h_2} h_2'(1) + \frac{2}{\langle n_3 \rangle} \frac{\partial^2 G_0(1,1)}{\partial h_3 \partial h_3} h_3'(1) \tag{4.31b}$$

The derivatives of G_0 are readily calculated as $\partial_{h_2 h_2} G_0 = \langle n_2^2 \rangle - \langle n_2 \rangle$, $\partial_{h_3 h_3} G_0 = \langle n_3^2 \rangle - \langle n_3 \rangle$, and $\partial_{h_2 h_3} G_0 = \langle n_2 n_3 \rangle$. Inserting these, multiplying Eq 4.31a by $\langle n_2 \rangle$ and Eq 4.31b by $\langle n_3 \rangle$ we have

$$h_2'(1)(2\langle n_2 \rangle - \langle n_2^2 \rangle) = \langle n_2 \rangle + 2\langle n_2 n_3 \rangle h_3'(1) \tag{4.32a}$$

$$h_3'(1)(3\langle n_3 \rangle - 2\langle n_3^2 \rangle) = \langle n_3 \rangle + \langle n_2 n_3 \rangle h_2'(1) \tag{4.32b}$$

Multiplying Eq 4.32a by $(3\langle n_3 \rangle - 2\langle n_3^2 \rangle)$, inserting the right hand side of Eq 4.32b and isolating $h_2'(1)$, we find

$$h_2'(1) = \frac{\langle n_2 \rangle (3\langle n_3 \rangle - 2\langle n_3^2 \rangle) + \langle n_2 n_3 \rangle \langle n_3 \rangle}{(3\langle n_3 \rangle - 2\langle n_3^2 \rangle)(2\langle n_2 \rangle - \langle n_2^2 \rangle) - 2\langle n_2 n_3 \rangle^2} \tag{4.33}$$

A similar process can be applied to Eq 4.32b to obtain an expression for $h_3'(1)$

$$h_3'(1) = \frac{\langle n_3 \rangle (2\langle n_2 \rangle - \langle n_2^2 \rangle) + \langle n_2 n_3 \rangle \langle n_2 \rangle}{(3\langle n_3 \rangle - 2\langle n_3^2 \rangle)(2\langle n_2 \rangle - \langle n_2^2 \rangle) - 2\langle n_2 n_3 \rangle^2} \tag{4.34}$$

These equations can now be used to solve for the average component size in Eq 4.30.

In the general case, when the determinant vanishes, $\det(\boldsymbol{I} - \boldsymbol{\alpha}^{-1}\boldsymbol{H}\boldsymbol{\beta}) = 0$, the average component size diverges, signalling the onset of the giant component. The generalisation

of the Molloy-Reed condition that indicates when a GCC can be found in the network is

$$\det(\boldsymbol{I} - \boldsymbol{\alpha}^{-1}\boldsymbol{H}\boldsymbol{\beta}) \leq 0 \tag{4.35}$$

For instance, when the network contains only tree-like edges, then the determinant in Eq 4.35 yields the familiar Molloy-Reed criterion, $\langle n_2^2\rangle/\langle n_2\rangle - 2 = 0$. When the network consists of tree-like and triangular edges, Eq 4.35 reduces to

$$\left(\frac{\langle n_2^2\rangle - \langle n_2\rangle}{\langle n_2\rangle} - 1\right)\left(2\frac{\langle n_3^2\rangle - \langle n_3\rangle}{\langle n_3\rangle} - 1\right) \leq 2\frac{\langle n_2 n_3\rangle^2}{\langle n_2\rangle\langle n_3\rangle} \tag{4.36}$$

a result obtained by [40, 52]. We can also obtain this result from the divergence of Eq 4.30 (i.e. when the denominator in Eqs 4.32a and 4.32b is zero). This result shows that a GCC can be formed in three ways: over the entire set of edges, independent of their topology; or, in either the triangles or the independent edges if the average number of independent edges or triangles vanishes, respectively. Similar findings occur for random graphs that have larger cliques [32]. For networks composed of a single clique-type of size ν, the Molloy-Reed criterion is given by

$$\left((\nu-1)\frac{\langle n_\nu^2\rangle}{\langle n_\nu\rangle} - \nu\right) \leq 0 \tag{4.37}$$

4.3 Single topology networks

The generating function $h_0(z)$ contains a lot of valuable information regarding the structure of the network. The coefficients, π_s, are the probability that a vertex selected at random belongs to a component of size s. Unfortunately, we cannot evaluate $h_0(z)$ directly; however, the coefficients of the generating function can, in some cases, be found.

In this section we examine how to extract the coefficients of $h_0(z)$ for random graphs that are constructed from a single clique topology and therefore, build the distribution of finite component sizes. The GCM construction process gives rise to components that are comprised of edge-disjoint cliques of a given size, see Fig 4.2. Due to the locally tree-like property of these components, the accidental joining of two motifs has a vanishingly small probability. The means that the size distribution of the finite components of GCM networks composed of η-cliques larger than the 2-clique (ordinary edges) is only non-zero at values of s given by

$$s = l(\eta - 1) + 1 \tag{4.38}$$

where l is the number of η-cliques in the component. In this case, the system of equations given by Eqs 4.1 is reduced to

$$h_0(z) = zG_0(h_\eta^{\eta-1}(z)), \qquad h_\eta(z) = zG_{1,\eta}(h_\eta^{\eta-1}(z)) \tag{4.39}$$

The smallest component possible is an isolated η-clique and therefore the generating function of the component sizes, $h_0(z)$, is of leading order of at least z. This means that $h_0(z)$ contains an overall factor of z which can be divided out. Differentiating $h_0(z)/z$, we can write the probability of belonging to a cluster of size s by extracting the $(s-1)$-th

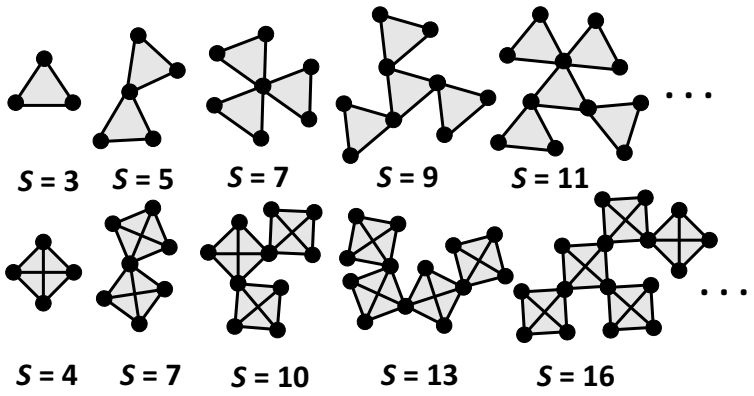


Figure 4.2: The small components for GCM networks consisting entirely of 3-clique (top) and 4-clique (bottom) motifs. Only values of $s = l(\eta - 1) + 1$, for integer $l = 1, 2, 3, \ldots$ and clique size η are permissible under the locally tree-like GCM construction constraints.

coefficient of z in $h_0(z)/z$. We find

$$\pi_s = \frac{1}{(s-1)!}\left[\frac{d^{s-1}}{dz^{s-1}}\left(\frac{h_0}{z}\right)\right]_{z=0} \tag{4.40}$$

$$= \frac{1}{(s-1)!}\left[\frac{d^{s-2}}{dz^{s-2}}\left(\frac{d}{dz}G_0(h_\eta^{\eta-1})\right)\right]_{z=0} \tag{4.41}$$

The innermost derivative is evaluated as

$$\frac{d}{dz}G_0(h_\eta^{\eta-1}) = (\eta-1)\langle n_\eta\rangle G_{1,\eta}(h_\eta^{\eta-1})h_\eta^{\eta-2}\frac{dh_\eta}{dz} \tag{4.42}$$

and so, we obtain

$$\pi_s = \frac{(\eta-1)\langle n_\eta\rangle}{(s-1)!}\left[\frac{d^{s-2}}{dz^{s-2}}\left(G_{1,\eta}(h_\eta^{\eta-1})h_\eta^{\eta-2}\frac{dh_\eta}{dz}\right)\right] \tag{4.43}$$

To proceed with the derivative, we must find a way to eliminate the requirement to directly evaluate h_η. This can be achieved by applying the Cauchy formula to the derivatives. In general, consider a holomorphic generating function $F(z_1,\ldots,z_d) = \sum_{\boldsymbol{r}} a_{\boldsymbol{r}}\boldsymbol{Z}^{\boldsymbol{r}}$ in d variables, with $\boldsymbol{r}$ being a d-tuple of integers and $\boldsymbol{Z}^{\boldsymbol{r}} = Z_1^{r_1}\cdots Z_d^{r_d} \in \mathbb{C}^d$. To recover the coefficients $\{a_{\boldsymbol{r}}\}$ from F we employ the Cauchy integral formula

$$a_{\boldsymbol{r}} = \frac{\boldsymbol{r}!}{(2\pi i)^d}\int_T \boldsymbol{Z}^{-\boldsymbol{r}}F(\boldsymbol{z})\frac{d\boldsymbol{Z}}{\boldsymbol{Z}} \tag{4.44}$$

where T is the torus comprising closed disks about the origin of each coordinate, $d\boldsymbol{Z}/\boldsymbol{Z}$ is $(Z_1\cdots Z_d)^{-1}$ times the holomorphic volume form $dz_1\wedge\cdots\wedge dz_d$ and $\boldsymbol{r}! = r_1!\cdots r_d!$.

For our case, we have

$$\frac{d^{s-2}}{dz^{s-2}}\left(G_{1,\eta}(h_\eta^{\eta-1})h_\eta^{\eta-2}\frac{dh_\eta}{dz}\right) = \frac{(s-2)!}{2\pi i}\oint \frac{1}{z^{s-2+1}}G_{1,\eta}(h_\eta^{\eta-1})h_\eta^{\eta-2}h'_\eta dz \tag{4.45}$$

$$= \frac{(s-2)!}{2\pi i}\oint \frac{[G_{1,\eta}(h_\eta^{\eta-1})]^s}{[h_\eta]^{s-1-\eta+2}}dh_\eta \tag{4.46}$$

$$= \frac{(s-2)!}{(s-\eta)!}\frac{d^{s-\eta}}{dz^{s-\eta}}\left[G_{1,\eta}(z^{\eta-1})\right]^s \tag{4.47}$$

where we inverted Eq 4.39 to consider z as a function of h_η and $G_{1,\eta}$, before applying the Cauchy formula once more. Therefore, the s-th coefficient of $h_0(z)$ is given by

$$\pi_s = \frac{(\eta-1)\langle n_\eta\rangle}{(s-1)!}\frac{(s-2)!}{(s-\eta)!}\frac{d^{s-\eta}}{dz^{s-\eta}}\left[G_{1,\eta}(z^{\eta-1})\right]^s \tag{4.48}$$

where the derivatives are to be evaluated at $z=0$. Importantly, this expression removes the need to evaluate h_η in order to obtain π_s. The only exception to this expression is when $s=1$, for which Eq 4.48 incorrectly yields yields $\pi_1 = 0/0$. However, since the only way to belong to a component of size 1 is to have no connections to any other vertices, the probability π_1 is trivially equal to the probability of having degree zero

$$\pi_1 = p_0 \tag{4.49}$$

For ordinary edges, $\eta = 2$, and we obtain Newman's result [51]

$$\pi_s = \frac{\langle k\rangle}{(s-1)!}\frac{d^{s-2}}{dz^{s-2}}\left[G_1(z)\right]^s \tag{4.50}$$

where $\langle k\rangle = \langle n_2\rangle$ is the average degree. For 3-cliques, $\eta = 3$ and we have

$$\pi_s = \frac{2\langle n_3\rangle}{(s-1)!}\frac{(s-2)!}{(s-3)!}\frac{d^{s-3}}{dz^{s-3}}\left[G_1(z^2)\right]^s \tag{4.51}$$

As an example, consider the finite components of a graph that is composed of edge-disjoint η-cliques. The distribution of the number of cliques a vertex belongs to, n_η, is Poisson distributed with mean $\langle n_\eta\rangle$. We have

$$[G_0(z^{\eta-1})]^s = e^{s\langle n_\eta\rangle z^{\eta-1}}e^{-s\langle n_\eta\rangle} \tag{4.52}$$

We will remove the constant $e^{-s\langle n_\eta\rangle}$ for now and also set $a = s\langle n_\eta\rangle$ and $\eta - 1 = m$ for ease. Consider the Taylor series of z with a small parameter $t \to 0$.

$$e^{a(z+t)^m} = \sum_{n=0}^{\infty}\lim_{t\to 0}\frac{d^n}{dt^n}e^{a(z+t)^m}\frac{t^n}{n!} \tag{4.53}$$

This is a common trick in combinatorics to find the n-th derivative: express a series expansion as the exponential generating function before applying a wealth of tools to find the n-th coefficient in closed-form. The n-th coefficient of this exponential generating

function is the n-th derivative we seek. We have

$$e^{a(z+t)^m} = \sum_{k=0}^{\infty} \frac{(a(z+t)^m)^k}{k!} \tag{4.54}$$

$$= \sum_{k=0}^{\infty} \frac{a^k}{k!} \sum_{j=0}^{mk} \binom{mk}{j} z^j t^{mk-j} \tag{4.55}$$

$$= \sum_{k=0}^{\infty} \frac{a^k}{k!} z^{mk} \sum_{j=0}^{mk-1} \binom{mk-1}{j} z^j t^{mk-1-j} \tag{4.56}$$

$$= \sum_{k=0}^{\infty} \frac{a^k}{k!} z^{mk} \sum_{j=0}^{mk-1} \sum_{r=0}^{j} \binom{mk}{r} \binom{-1}{j-r} z^j t^{mk-1-j} \tag{4.57}$$

where in the last step we have used Vandermonde's identity to expand the binomial coefficient

$$\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k} \binom{n}{r-k} \tag{4.58}$$

Consider the terminal negative binomial coefficient in more detail. In general we can write

$$\binom{-r}{k} = \frac{1}{k!} \prod_{i=0}^{k-1} (-r-i) \tag{4.59}$$

$$= \frac{(-1)^k}{k!} \prod_{i=0}^{k-1} (r+i) \tag{4.60}$$

$$= \frac{(-1)^k}{k!} \frac{(r+k-1)}{(r-1)!} \tag{4.61}$$

$$= (-1)^k \binom{r+k-1}{k} \tag{4.62}$$

When $r = 1$ and $k = j - r$ we have

$$\binom{-1}{j-r} = (-1)^{j-r} \tag{4.63}$$

Inserting this back into the main expression we have

$$e^{a(z+t)^m} = \sum_{k=0}^{\infty} \frac{(az^m)^k}{k!} \sum_{j=0}^{mk-1} \sum_{r=0}^{j} \frac{(mk)!}{r!(mk-r)!} (-1)^{j-r} z^j t^{mk-1-j} \tag{4.64}$$

All that remains is to extract the n-th coefficient of this series, which we achieve by re-indexing the summation and comparing to Eq 4.53 to arrive at

$$\frac{d^n}{dz^n} e^{az^m} = e^{az^m} z^{-n} \sum_{k=0}^{n} \sum_{j=0}^{k} \frac{(-1)^j (az^m)^k (1 - jm + km - n)_n}{j!(k-j)!} \tag{4.65}$$

where $(x)_n$ is the Pochhammer symbol

$$(x)_n = \prod_{k=0}^{n-1}(x-k) \tag{4.66}$$

This is equivalent to an exponential Riordan array, $B_m = [1,(1+x)^m+1]$. A Riordan array is a lower triangular matrix whose elements satisfy a recurrence relation. An exponential Riordan array is a matrix that is defined by a pair of generating functions $f(x)$ and $g(x)$ such that the k-th column has exponential generating function $g(x)f(x)^k/k!$, and satisfies criteria to be defined as a member of a Riordan group. Numerically, all this means is that we can use a recurrence relation to obtain the n-th derivative as a lower triangular matrix

$$\frac{d^n}{dx^n}e^{ax^m} = \left(\sum_{k\in\mathbb{Z}} b_{n,k;m}(ax)^{km-n}\right)e^{ax^m} \tag{4.67}$$

It remains to insert these values into Eq 4.48 to obtain π_s. For triangles the n-th derivative of e^{az^2} yields a closed-form expression in terms of the Hermite polynomials; which, when evaluated at $z=0$ produces the Hermite numbers, H_n. The Hermite numbers are only non-zero when $s-3$ is even; therefore, π_s only takes non-zero values for odd s, as predicted by Eq 4.38. We confirm the accuracy of our expressions by comparing the results against Monte Carlo simulation in Fig 4.3.

We remark that the coefficients of $h_0(z)$ can be found by application of the Cauchy formula as was performed in section 4.3. The Cauchy formula approach to extract the coefficients of a generating function is a particular of the *Lagrange inversion theorem* [15, 29]. The theorem is as follows. Given $\phi(x)$ and $\rho(x)$ are formal power series such that $\phi(x) = x\rho(\phi(x))$, then for formal series $f(x)$ the coefficient of $f(\phi(x))$ at x^n is given

$$[x^n]f(\phi(x)) = \frac{1}{n}[x^{n-1}]f'(x)\rho^n(x)$$

where $[x^n]f(x)$ is shorthand notation for the coefficient of $f(x)$ at x^n. In the network science literature we substitute $f(z) = G_0(z)$, $\phi(z) = h_1(z)$ and $\rho(z) = G_1(x)$ to obtain Newman's result [51] directly from the theorem.

4.4 Finite components of arbitrary GCM networks

In section 4.3 we examined GCM networks that are composed of a single clique topology. In this section we generalise this expression to account for random graphs that are composed of an arbitrary number of edge-disjoint cliques. The evaluation of π_s proceeds in the same manner as previously, until the derivative of G_0 is calculated

$$\pi_s = \frac{1}{(s-1)!}\left[\frac{d^{s-2}}{dz^{s-2}}\left(\frac{d}{dz}G_0(h_2(z),h_3^2(z),\ldots,h_m^{m-1}(z))\right)\right]_{z=0} \tag{4.68}$$

where the derivative is evaluated as

$$\frac{d}{dz}G_0(h_2(z),h_3^2(z),\ldots,h_m^{m-1}(z)) = \sum_{\eta}(\eta-1)\langle n_\eta\rangle G_{1,\eta}(h_\eta^{\eta-1})h_\eta^{\eta-2}\frac{dh_\eta}{dz} \tag{4.69}$$

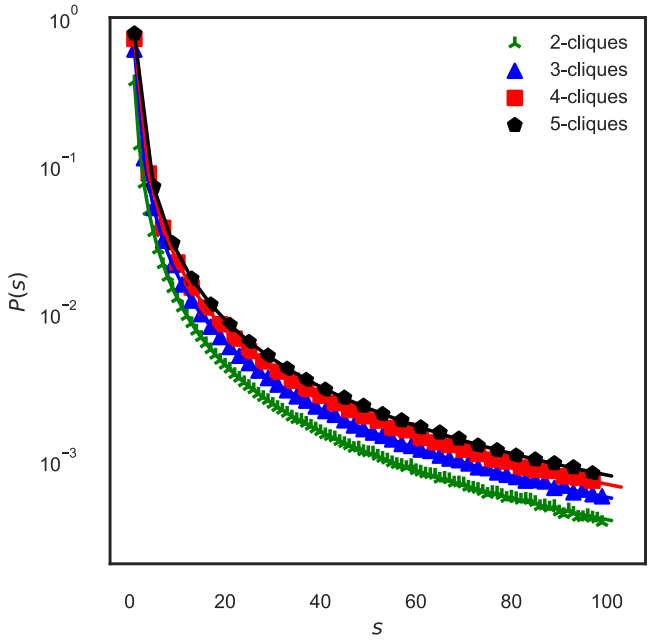


Figure 4.3: The small components for GCM networks consisting entirely of edge-disjoint m-cliques. Scatter points are the average of 250 repeats of Monte Carlo simulation on networks with $2e5$ vertices with clique membership Poisson distributed with fixed first moment $(m-1)\langle k_m\rangle = 1$ across all experiments. The curves are the theoretical results of Eq 4.48, evaluating the required derivatives of the Poisson generating function using the Riordan array recurrence relation given in 4.67.

By the linearity of the derivative operator, the Cauchy formula now acts on each term in the sum and so we find

$$\pi_s = \sum_{\eta} \frac{(\eta-1)\langle n_\eta\rangle}{(s-1)!}\left[\frac{d^{s-2}}{dz^{s-2}}\left(G_{1,\eta}(h_\eta^{\eta-1})h_\eta^{\eta-2}\frac{dh_\eta}{dz}\right)\right] \tag{4.70}$$

Expanding this expression with the Cauchy formula, inverting $h_\eta(z) = zG_{1,\eta}(h_2,\ldots,h_m^{m-1})$ for z and considering each $h_\eta(z)$ as an independent holomorphic variable (so we can undo the Cauchy formula) we have

$$\pi_s = \sum_{\eta} \frac{(\eta-1)\langle n_\eta\rangle}{(s-1)!}\frac{(s-2)!}{(s-\eta)!}\frac{d^{s-\eta}}{dz^{s-\eta}}\left[G_{1,\eta}(z^{\eta-1})\right]^s \tag{4.71}$$

This expression is simply the sum of the components along each edge topology in Eq 4.48. We use this expression to investigate the effect of clustering on the finite components of mixed 2- and 3-clique networks in Fig 4.4. We notice that clustering tends to increase π_s for a given s; however, we will not draw premature conclusions until further studies have been conducted that also control the overall degree assortativity.

An interesting observation is that the distribution of π_s can oscillate at small s before it converges to its asymptote as s grows, see $(n_2,n_3) = (0.25, 0.625)$ in Fig 4.4. We believe this is due to the mixing of the components that can be made from 2-cliques and those that

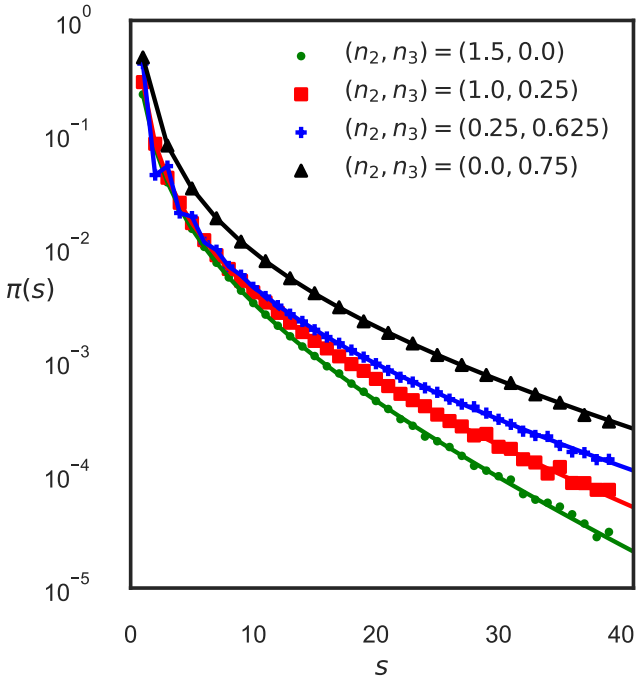


Figure 4.4: The small components for GCM networks consisting of edge-disjoint 2- and 3-cliques. Scatter points are the average of 20 repeats of Monte Carlo simulation on networks with $2e5$ vertices with clique membership Poisson distributed with fixed first moment $\langle n_2 \rangle + 2\langle n_3 \rangle = 1.5$ across all experiments. The curves are the theoretical results of Eq 4.71.

can be made from 3-cliques. For instance, triangles cannot contribute to components of size $s = 2$; however, components of size $s = 3$ can be made either 2- or 3-clique motifs. Therefore, the number of available vertices to contribute to each value of s is different.

4.5 Bond percolation threshold

We now turn our attention to the location of the critical point for the formation of a GCC among networks comprised entirely of η-cliques during bond percolation. To obtain the percolation properties of the network, we have to evaluate the derivative of $g_\eta^{\eta-1}$ with respect to u. This derivative is found to be

$$\frac{\partial g_\eta^{\eta-1}}{\partial u} = \sum_{r=0}^{\eta-1} \binom{\eta-1}{r} (\eta - r - 1) \sum_{j=0}^{E(\eta-r)} q_{\eta-r, X_{\eta-r,j}} \times (uT)^{\eta-r-2} T^{E(\eta-r)-j} (1-T)^{\omega(r)} \tag{4.72}$$

The percolation threshold is then obtained by evaluating the derivative at $u = 1$, and following a similar analysis to the tree-like topology we obtain

$$\left.\frac{\partial g_\eta^{\eta-1}}{\partial u}\right|_{u=1} \frac{\langle n_\eta^2 - n_\eta \rangle}{\langle n_\eta \rangle} = 1 \tag{4.73}$$

where $\langle n_\eta \rangle$ is the mean number of η-cliques a vertex belongs to. For example, the derivative for 3-cliques is found to be

$$\frac{\partial g_3^2}{\partial u} = 2T(1-T)^2 + 6uT^2(1-T) + 2uT^3 \tag{4.74}$$

Evaluated at $u = 1$ and inserted into Eq 4.73 we have

$$2(T^2 + T - T^3)\frac{\langle n_3^2 - n_3 \rangle}{\langle n_3 \rangle} - 1 = 0 \tag{4.75}$$

For networks in which a vertex's membership in a given clique size is Poisson distributed, we can reduce $\langle n_\eta^2 - n_\eta \rangle / \langle n_\eta \rangle$ to simply $\langle n_\eta \rangle$. Further, factorising the pre-factor in T we have $2T(1 + T - T^2)\langle n_3 \rangle - 1 = 0$. Using the discriminant, this cubic expression is reducible in T into the quadratic form whose roots yield the critical transmissibilities of the model, and hence, the critical point occurs at

$$T^*_{\text{Poisson}} = -1 + 2\sqrt{1 + \frac{1}{\langle n_3 \rangle}} \tag{4.76}$$

We repeat the calculation for the 4-clique to obtain the following polynomial

$$\left.\frac{\partial g_4^3}{\partial u}\right|_{u=1} = 3T(-2T^5 + 7T^4 - 7T^3 + 2T + 1) \tag{4.77}$$

The Galois group of the quintic part is the symmetric group, S_5, which means that a root cannot be found. It is unlikely that percolation properties of larger cliques can be resolved analytically due to the Abel-Ruffini theorem, which states that there is no solution in radicals to general polynomial equations of degree five or higher with arbitrary coefficients.

4.6 Chapter summary

In this chapter we have investigated the properties of the finite (non-giant) components of clustered networks. We first derived an expression for $h_0(z)$ from first principles before using this expression to find the mean component size for GCM networks. We showed that the expression diverged at the critical point of the model when an infinite cluster first appears. We then turned our attention to the analytical form of the distribution of component sizes for random networks; restricting our focus to networks composed of a single clique topology. We found that the component size distribution took non-zero values only at specific sizes and supported this by a simple counting argument. We then generalised this to GCM networks with multiple clique topologies. In all cases we found excellent agreement between the analytical expressions and Monte Carlo simulation. The

bond percolation threshold of clique networks was then derived and it was postulated that the roots of the resulting polynomials could not be deduced, in this manner, for cliques larger than $\eta = 5$.

Upon first inspection, for Poisson distributed motifs with fixed first moment, our results show that larger clique sizes increase the component probability distribution π_s. Further, increased clustering coefficient for fixed clique size mixed-topology GCM networks also increases π_s. This is supported by the reduction in the percolation threshold, evidenced from the divergence of the mean component size for these networks. However, it is well-known that, for Poisson distributed GCM networks, clustering also increases the degree assortativity; therefore, more experiments are required, using the degree-δ model, to understand the role of assortativity.

CHAPTER FIVE

FINDING g_η FOR ARBITRARY SUBGRAPHS

In the preceding chapters, we have examined methods to construct the GCC for random graphs that are comprised of clique subgraphs. Cliques have perfect symmetry. As we saw in section 3.2, a closed-form, exact expression for $g_\eta^{\eta-1}$ for all sizes of clique can be written. In this chapter we discuss subgraphs that are not cliques and thus, in general, do not have perfect symmetry. Of course, there are many other motifs that are symmetric; for instance chordless cycles, or k-regular motifs, see Fig 5.1.

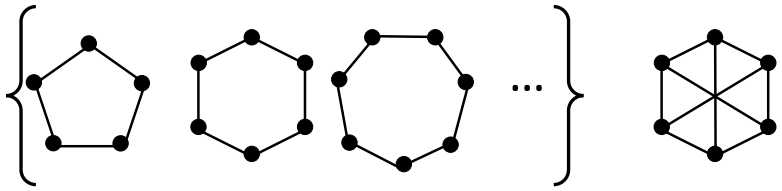


Figure 5.1: A series of chordless cycles (curled brackets) with increasing size; and, an example of a k-regular subgraph on 6 vertices (right). The 3-clique also belongs to both of these sets.

We will examine the g_η expression for these special cases and discover a hidden complexity with these motifs. We will show that the important property that cliques hold is not symmetry, rather, it is the induction of a clique subgraph of size $\eta - r$ when r vertices are removed from an η-clique as well as the fact that all possible edges between each vertex pairs are present. These properties motivate the recursion relation derived by Gilbert [17] and later Newman [47]. We will formalise the exact enumeration of g_η for arbitrary motifs, which is NP-hard, before introducing a counting methodology based on the closed-form expression for cliques, see chapter 3.2. In the final part of this chapter we then consider a different approach, based on the SIR equivalence, that approximates g_η by enumerating the non-self-intersecting walks in a motif.

5.1 Chordless cycles

Perhaps the simplest set of subgraphs to consider is the set of chordless cycles of increasing vertex count. In the ordinary configuration model, chordless cycles of length $O(\log N)$ are formed accidentally during the construction process [56]; therefore, the resulting graphs are only guaranteed to be *locally* tree-like. In this section we expressly incorporate chordless cycles into the GCM to investigate their properties under bond percolation. It happens, due to their simplicity, that a closed-form expression for the probability that a focal vertex fails to be connected to the GCC despite its membership in an η-cycle, g_η^2, can be obtained [32]. The exponent of g_η^2 is always 2 because all vertices in a chordless η-cycle have degree 2. To see this, consider a chordless η-cycle with all of its edges occupied, and hence, all of its vertices in the unconnected state. A single edge can be unoccupied before a vertex is removed and so the coefficient of $u^{\eta-1}$ has two terms

$$P(\eta,0) = \sum_{j=0}^{1} \binom{\eta}{j} (uT)^{\eta-1} T^{1-j} (1-T)^j \tag{5.1}$$

Subsequent removal of an edge will isolate a vertex, and so, all states belonging to the term proportional to $u^{\eta-1}$ have been exhausted.

We now examine the case that a single neighbour vertex within the cycle belongs to the GCC. For the rest of the cycle to remain unattached to the GCC, both of the edges that connect to this vertex must be unoccupied. There are $\eta-1$ vertices to choose from and so, we have

$$P(\eta,1) = \binom{\eta-1}{1} (uT)^{\eta-1-1} (1-T)^2 \tag{5.2}$$

No edges can be removed from this motif without further isolation of a vertex, indicating that the coefficient of $u^{\eta-2}$ is complete. Considering the next term, $u^{\eta-3}$, we find

$$P(\eta,2) = \binom{\eta-2}{1} (uT)^{\eta-1-2} (1-T)^2 \tag{5.3}$$

More generally for the removal of r vertices we have

$$\sum_{r=1}^{\eta-1} P(\eta,r) = \sum_{r=1}^{\eta-1} \binom{\eta-r}{1} (uT)^{\eta-1-r} (1-T)^{(r+1)-(r-1)} \tag{5.4}$$

where the power of $(1-T)$ is the number of edges that no longer have an occupied path to the focal vertex, $(r+1)$, minus the number within the removed component that we cannot specify the state of $(r-1)$; this evaluates to $(r+1)-(r-1)=2$; since, only the connections to the focal vertex are required to be unoccupied. The summation over r extends to $r=\eta-1$, such that the power of u on the final term is u^0, which corresponds to an isolated focal vertex, with both of its edges unoccupied. The final expression for the total probability that a focal vertex remains in the RG despite its membership in an η cycle, g_η^2, requires the combination of Eq 5.1 and Eq 5.4 into a single expression. The most straightforward way to achieve this is simply to sum the independent terms

$$g_\eta^2 = P(\eta,0) + \sum_{r=1}^{\eta-1} P(\eta,r) \tag{5.5}$$

where

$$P(\eta,0)+\sum_{r=1}^{\eta-1}P(\eta,r)=\sum_{j=0}^{1}\binom{\eta}{j}(uT)^{\eta-1}T^{1-j}(1-T)^{j} + \sum_{r=1}^{\eta-1}\binom{\eta-r}{1}(uT)^{\eta-1-r}(1-T)^{(r+1)-(r-1)} \tag{5.6}$$

With an exact closed-form expression for g_η^2 for chordless cycles Mann *et al* [32] showed that cycles of increasing size behave increasingly similar to the 2-clique approximation, see Fig 5.2. In other words, the correlations induced by the presence of a closed loop is largest for triangles and becomes reduced as the path length of the loop grows. This is the foundation for the success of the tree-like approximation: the properties induced by short cycles disappear as the cycle length approaches $\log N \to \infty$.

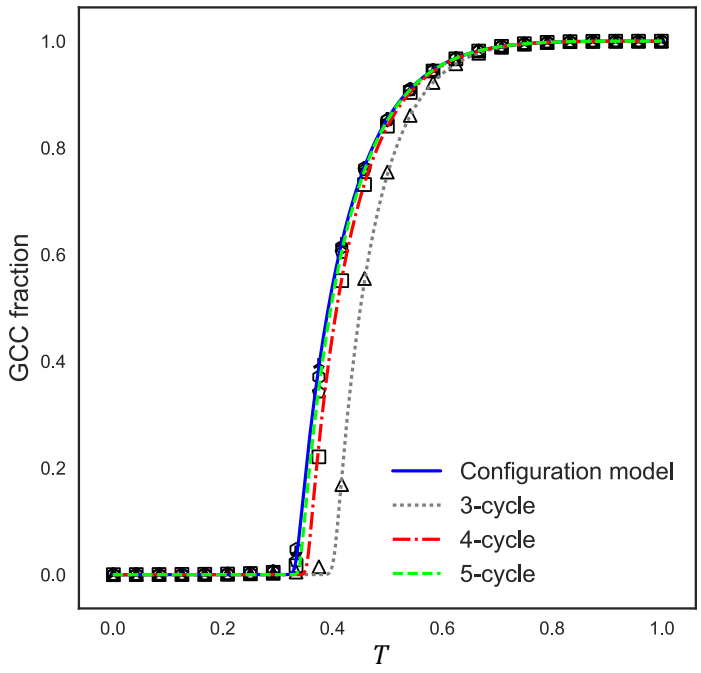


Figure 5.2: The size of the GCC for random graphs composed of chordless cycles where each vertex has a fixed degree $k = 4$ and is therefore a member of two cycles. Intuitively, larger cycles behave increasingly tree-like. Scatter points are from Monte Carlo simulation whilst plotted lines are theoretical results. Figure reproduced from [32].

5.2 k-regular subgraphs

In this section we examine the possibility of a closed-form percolation expression for subgraphs whose vertices are degree equivalent to one another (k-regular), but that are not

either cliques or chordless cycles. We demonstrate that the expression for these subgraphs is not readily obtained in closed-form for all sizes, unlike the expressions for chordless cycles and cliques. Any formula that can be obtained for a class of subgraphs appears to be distinct from the expressions for other, seemingly related, motifs. An exact expression for these motifs therefore relies upon an exhaustive enumeration of states once more. For this purpose, consider a subgraph comprising of $\eta \geq 6 \in 2\mathbb{N}$ degree equivalent vertices in which each vertex has degree 3, see Fig 5.1 (right). (Note the motif cannot be formed for odd η.) Application of the enumeration scheme developed in section 3.2 to obtain the probability that a particular focal vertex does not become attached to the GCC through its role in this motif proceeds as follows. The subgraph can lose up to $\eta/2+1$ edges before a vertex is isolated, the probability that $j \in [0, \eta/2+1]$ are removed is given by

$$P(j \mid \eta, 0) = q'_{\eta,\eta+\eta/2-j} u^{\eta-1} T^{\eta} T^{\eta/2-j} (1-T)^{j} \tag{5.7}$$

where $q'_{n,k}$ is the number of conencted graphs that can be made with n vertices and k edges that are also subgraphs of the original motif. Hence the total probability $P(\eta, 0)$ that we can still retain a connected graph despite the removal of edges is

$$P(\eta, 0) = \sum_{j=0}^{\eta/2+1} P(j \mid \eta, 0) \tag{5.8}$$

With any further edge removal, a vertex is pruned from the motif. The resulting cycle has $\eta - 3$ deg(3) and 3 deg(2) sites in the subgraph.

There are now $\eta/2-1$ interior edges and $\eta-2$ exterior edges remaining from the original set of edges. It happens that we can remove all of the remaining interior edges and proceed without vertex-isolation; however, we cannot remove any of the exterior edges. Thus, the total probability $P(\eta, 1)$ that describes the motif with one vertex removed is given by

$$\begin{aligned} P(\eta, 1) = (\eta - 1) \sum_{j=0}^{\eta/2-1} & q'_{\eta-1,\eta-2+\eta/2-1-j} u^{\eta-2} \\ & \times T^{\eta-2} T^{\eta/2-1-j} (1-T)^{j+3} \end{aligned} \tag{5.9}$$

At this point, the subgraph now contains mixed degree vertices. We must distinguish upon whether the vertex we now remove has degree 2 or degree 3 as removing either vertex will lead to different probabilities for successive counting. Further, considering Fig 5.3 where $\eta = 10$, supposing we had removed a degree 3 vertex from the original motif (left) to generate the middle motif, we must distinguish whether the neighbours of the currently considered deg(3) vertex are themselves deg(2) and deg(3) (Fig 5.3 blue vertex) or are both deg(3) (Fig 5.3 red vertex) as in each case, the resulting probabilities for further counting are non-equivalent; they are history dependent. And, whilst it is certainly theoretically possible to enumerate the combinations into a single expression, it seems unlikely that such a formula would be readily derived, or that it would be transferable to other related cycles that differ even by a single edge. Similar complexities arise for other k-regular motifs and this is the basis of the complexity in enumerating percolation formulas for arbitrary subgraphs [36].

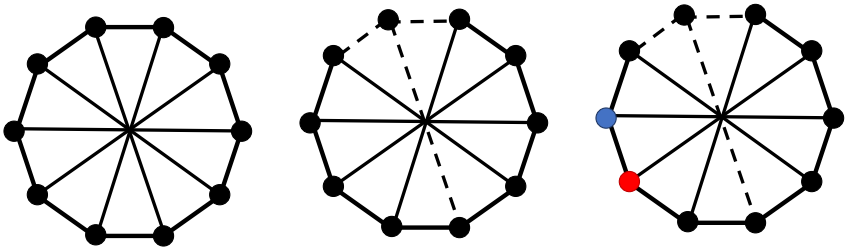


Figure 5.3: A k-regular cycle on 10 vertices where each vertex also has a chord (left). The removal of a vertex (dashed edges) results in $3 \times \deg(2)$ sites within the motif, in addition to the original $(\eta - 3) \times \deg(3)$ sites. Further counting depends on the degree of the focal vertex as well as the neighbours of the focal vertex. For instance, the blue and red vertices have neighbours of different degree. Both of these factors lead to different probabilities when enumerated, and so, complicate the counting procedure.

5.3 Arbitrary subgraphs

Across the preceding sections, all of the subgraphs that we have considered so far have been symmetric prior to percolation. In this section we relax this condition to consider subgraphs whose topology is arbitrary. A closed-form solution for subgraphs with arbitrary topology would render the percolation problem on random graphs exactly solved for all graphs. One could imagine considering the largest Hamiltonian cycle that could be formed from the graph as a subgraph for the model. It is speculated that this might capture the long range correlations between distant neighbours in networks with more accuracy than an edge-disjoint clique cover perhaps could. Consider the specific examples of subgraphs

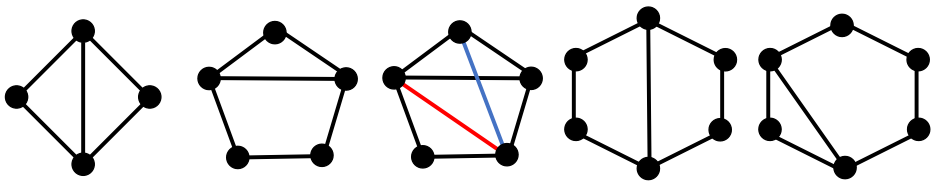


Figure 5.4: Subgraphs with arbitrary topology.

with arbitrary topology in 5.4. From left to right we have: a subgraph on 4 vertices with one chord; a 5-vertex chorded subgraph with a single chord, a 2-chord 5-vertex subgraph (where the choice of the two distinct options for placing the second chord are highlighted in blue and red); and finally, two distinct single chord subgraphs on 6 vertices. It is clear that as the cycles become larger, there are multiple locations to place a single additional chord, and in the case of a 6-vertex cycle, even the location of the first chord is not unique.

Karrer and Newman [25] introduced a framework for the elucidation of the percolation properties of graphs comprised of subgraphs with arbitrary topologies. Their model is

similar to the system of equations in Eq 2.47; where each clique topology has its own argument in the G_0 generating function and its own excess degree distribution. Now, an argument for each unique site within a subgraph must be created as well as an excess degree distribution. For instance, consider a network composed of the 4-vertex chorded cycle in Fig 5.4. There are two unique sites in the cycle: $2 \times \deg(2)$ sites and $2 \times \deg(3)$ sites. A vertex that is a member of more than one 4-cycle could be the $\deg(2)$ site in one of the cycles and the $\deg(3)$ site in the other, see Fig 5.5. Given this, the joint probability

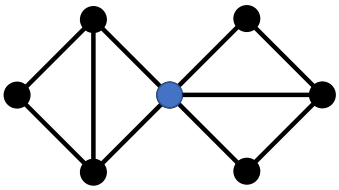


Figure 5.5: A focal vertex (blue) that is a member of two chorded 4-cycles as both the $\deg(2)$ site and the $\deg(3)$ site.

of choosing a vertex at random that is a member of i chorded 4-cycles as the $\deg(2)$ site and j chorded 4-cycles as the $\deg(3)$ site is p_{ij}; which is generated by

$$G_0(x_2,x_3) = \sum_{i=0}\sum_{j=0} p_{ij}x_2^i x_3^j \tag{5.10}$$

Each site has an excess degree distribution which are respectively generated as

$$G_{1,2}(x_2,x_3) = \frac{1}{\langle i\rangle}\frac{\partial G_0(x_2,x_3)}{\partial x_2} \tag{5.11}$$

$$G_{1,3}(x_2,x_3) = \frac{1}{\langle j\rangle}\frac{\partial G_0(x_2,x_3)}{\partial x_3} \tag{5.12}$$

where $\langle i\rangle$ is the number of $\deg(2)$ sites that the average vertex belongs to in the network whilst $\langle j\rangle$ is the average number of $\deg(3)$ sites.

The probability that a $\deg(2)$-site neighbour to a particular focal vertex does not belong to the GCC is u_2; with u_3 similarly defined for the $\deg(3)$ site. The probability that a focal vertex fails to be attached to the GCC despite its membership in a $\deg(3)$ site is $g_2^2(u_2,u_3)$; whilst, the probability that a focal vertex in a $\deg(3)$ site fails to be attached is $g_3^3(u_2,u_3)$. We notice that the g_η^η expressions are now coupled in both u variables. Once the g_η^η expressions have been evaluated (which we will discuss in a moment), the u_η values can be calculated as

$$u_\eta = G_{1,\eta}(g_2^2, g_3^3) \tag{5.13}$$

and the size of the GCC is given by

$$S = 1 - G_0(g_2^2, g_3^3) \tag{5.14}$$

Extrapolating this logic, and with a slight notational change, we now arrive at Karrer and Newman's model as follows. The joint probability distribution for the probability of randomly selecting a vertex that is a member of m topologically distinct cycles, each of

which has m_n different sites is

$$p_{1_1\ldots1_n\ldots m_1\ldots m_n} \tag{5.15}$$

which is generated as

$$G_0(x_{1_1},\ldots,x_{m_n}) = \sum_{k_{1_1}=0}^{\infty}\cdots\sum_{k_{m_n}=0}^{\infty} p_{1_1\ldots m_n} x_{1_1}^{k_{1_1}}\cdots x_{m_n}^{k_{m_n}} \tag{5.16}$$

The excess degree distribution for each edge-topology is defined as

$$G_{1,i_j} = \frac{1}{\langle k_{i_j}\rangle}\frac{\partial G_0}{\partial x_{i_j}} \tag{5.17}$$

The probability that a vertex in the i-th site of an η-cycle fails to belong to the GCC is u_{η_i}; which is generated as

$$u_{\eta_i} = G_{1,\eta_i}(g_{1_1},\ldots,g_{m_n}) \tag{5.18}$$

where we have dropped the exponent notation of g_{η_i}, the probability that a vertex in an η_i site fails to be attached to the GCC, for generality. The size of the GCC now follows from

$$S = 1 - G_0(g_{1_1},\ldots,g_{m_n}) \tag{5.19}$$

As with the other methods that are constructed in this way, the percolation threshold can be found by performing a linear stability analysis around the fixed point $u_{\eta_i} = 1$, $\forall\eta_i \in [\eta_1,\eta_n]$, $\forall\eta \in [1,m]$.

It remains now to calculate the g_{η_i} polynomial for each site of each subgraph that is included in the model; a process which is described by Karrer and Newman to be *exponentially slow* [25] and we recognise as NP-hard. Extending their work, we now outline our own counting method to enumerate the required combinations to find an exact expression for g_{η_i} for arbitrary motifs. Consider a motif η with η_N vertices and η_n distinct, labelled sites. We consider sites to be equivalent if the motif has either rotational or reflectional symmetry; otherwise if a symmetry operation cannot be performed, the sites are distinct. For example, the number of sites for each motif in Fig 5.4 (left to right) is given by: 2 for the 4 cycle, 3 for the single chorded 5-cycle, 3 for the red and 5 for the blue 2-chorded cycles, 2 and 4 for the 1-chord 6-cycles.

Let the probability that a vertex in site η_i fails to become attached to the GCC, despite its membership in the motif, be given by g_{η_i}. The probability that a vertex in a η_j-site is unattached to the GCC is given by u_{η_j}. Consider the set, $\{u_\eta\}$, that contains all $u_{\eta_j}, \forall j \in [1,\eta_n]$ where each element u_{η_j} is inserted as many times as it occurs in the motif. For instance, for the chorded 4-cycle the set would be $\{u_4\} = \{u_{4_2}, u_{4_2}, u_{4_3}, u_{4_3}\}$; since, there are two $\deg(2)$ sites and two $\deg(3)$ sites.

The form of g_{η_i} is a polynomial in all combinations of the elements of $\{u_\eta\}$, including the empty set, excluding a single copy of u_{η_i}, which is currently being considered as the focal vertex. For instance, g_{4_3} is a polynomial in all combinations of $\{u_{4_2}, u_{4_2}, u_{4_3}\}$ such that

$$\begin{aligned} g_{4_3} = {} & C_{\{4_2,4_2,4_3\}} + C_{\{4_2,4_2\}}u_{4_3} + 2C_{\{4_2,4_3\}}u_{4_2} \\ & + 2C_{\{4_2\}}u_{4_2}u_{4_3} + C_{\{4_3\}}(u_{4_2})^2 + C_{\{\emptyset\}}(u_{4_2})^2 u_{4_3} \end{aligned} \tag{5.20}$$

where C_λ are the as yet undetermined coefficients; the subscript $\lambda = \{\cdots\}$ indicates the

identities of vertices in the motif that belong to the GCC, which are subsets of $\{u_\eta\}$. Each coefficient is itself a polynomial of the form

$$C_\lambda = \sum_{r=0}^{E} q'_{n,k-r} T^{k-r} (1-T)^{W(k)+r} \tag{5.21}$$

where $n = N - \mathrm{card}(\lambda)$ and k are the number of vertices and edges in the motif at the current starting point, respectively. The number of edges, k is given by the maximum number of edges that can be occupied without attaching vertices in the RG to the GCC. The number of connected graphs that can be made from n vertices and k edges that are also subgraphs of the original motif is given by $q'_{n,k}$. Index r accounts for the number of edges, up to a maximum of E, that can be removed whilst still being able to make a connected graph of size equal to $N - \mathrm{card}(\lambda)$. Note, E is not simply $n-1$; since, we impose that the graphs be subgraphs of the original motif; it is determined by the topology of the specific motif. Finally, $W(k)$ is a function that accounts for the number of unoccupied edges required to ensure that vertices in the RG do not have an occupied path to vertices in the GCC. All of these quantities are dependent on the topology of the original motif as well as the set of removed vertices, $\{\lambda\}$.

Consider once more the calculation of g_{4_3}; with reference to the polynomial in Eq 5.20 we now examine the coefficient $C_{\{\emptyset\}}$ that accounts for the case where all vertices belong to the RG. The polynomial is given by

$$C_{\{\emptyset\}} (u_{4_2})^2 u_{4_3} = T^3 \left(q'_{4,5} T^2 + q'_{4,4} T(1-T) + q'_{4,3} (1-T)^2 \right) (u_{4_2})^2 u_{4_3} \tag{5.22}$$

The number of connected graphs that are subgraphs of the original motif can be calculated by enumerating the number of ways that the edges can be removed less those combinations that isolate a vertex; we have

$$q'_{4,5} = \binom{5}{0}, \qquad q'_{4,4} = \binom{5}{1}, \qquad q'_{4,3} = \binom{5}{2} - 2, \tag{5.23}$$

Subsequent edge removal would isolate a vertex, and so, we have enumerated all graphs. Next, assuming that the 4_3 site belongs to the GCC we have

$$C_{\{4_3\}} (u_{4_2})^2 = T^2 (1-T)^3 (u_{4_2})^2 \tag{5.24}$$

If one of the 4_2 sites belong to the GCC instead $C_{\{4_2\}}$ can be constructed as

$$2C_{\{4_2\}} u_{4_2} u_{4_3} = 2 \left(T^3 (1-T)^2 + \binom{3}{1} T^2 (1-T)^3 \right) u_{4_2} u_{4_3} \tag{5.25}$$

The counting process can be continued for each term in the polynomial expression for g_{4_3} where we find

$$2C_{\{4_2,4_3\}} u_{4_2} = 2T(1-T)^3 u_{4_2} \tag{5.26a}$$

$$C_{\{4_2,4_2\}} u_{4_3} = T(1-T)^4 u_{4_3} \tag{5.26b}$$

$$C_{\{4_2,4_2,4_3\}} = (1-T)^3 \tag{5.26c}$$

Having derived an exact formula g_{4_3}, one must now consider the 4_2 site as the focal vertex

and derive the probability that it remains unattached to the GCC, which we do not perform here. It is clear that the expressions that are obtained by this enumeration method are not transferable to other motifs, even if they differ by the presence, or indeed the placing of a single chord across the substrate motif.

We also observe that for cliques, the number of connected graphs $q_{n,k}$ is a known result from graph theory; whilst, in general the number of connected subgraphs $q'_{n,k}$ depends on the motif topology. When all connections between each vertex pair are possible in the substrate motif, $q'_{n,k} = q_{n,k}$.

5.4 An approximate method based on non-self intersecting walks

In section 5.3 we discussed Karrer and Newman's formulation [25] for deriving the probability, g_{η_i}, that a vertex in an η_i site is attached to the RG for arbitrary subgraphs. [1] Mann *et al* developed a method that yields an approximate formula for the g_{η_i} probability for arbitrary subgraphs based on enumerating the non-self intersecting walks within a motif.

The approximation is based on Miller's exact logic for triangles from [39] (see Eq A.3 from appendix A). In essence, for cliques, the enumeration of $g_\eta^{\eta-1}$ considers a vertex as the focal vertex and then, for each neighbour in the clique, considers the probability associated to each non-self-intersecting walk from the neighbour to the focal vertex as a possible connection pathway. Other vertices that are in the connection pathway also belong to the GCC; whilst vertices outside of the connection pathway belong to the RG. As with the exact formula, edges that connect vertices in the RG to vertices in the GCC must fail to be occupied. Thus, we must enumerate all walks back to the focal vertex starting from each neighbour. We will spare the reader the derivation of the expression and instead will consider a motivating example using the 4-clique. The key to understanding the formulation is that walks of a given length in a clique have equal probability of occurring. Therefore, we must count all walks of a given length through the clique from all potential source vertices to the focal vertex and then enumerate the probability of this walk. We find that the probability that a vertex involved in a 4-clique belonging to a finite-sized component during bond percolation is given by

$$\begin{aligned} g_4^3 =& (1-T+u_4T)^3 - 6(1-u_4)u_4T^2(1-T)(1-u_4T^2)^2(1-(1-u_4)T) \\ & - 6(1-u_4)u_4^2T^3(1-T)^3 \end{aligned} \tag{5.27}$$

The rationalisation of this expression is quite simple and can be read from left to right as follows. Consider a 4-clique and choose a vertex to be the focal vertex. The first cubic term relates to the failure of the three direct-contact vertices to connect the focal vertex to the GCC. These are 0-hop walks as they concern the direct linkage to the focal vertex. Labeling the vertices according to Fig 5.6 (left) we notice that if vertex 1 fails to connect the focal vertex directly, it can still connect it through edges in the clique. There are two

[1] Retrospectively, it is clear that the counting scheme that we formalised in section 5.3 can readily be converted to an algorithm that yields a symbolic expression for the polynomials we require. However, Mann *et al* were not aware of Karrer and Newman's paper [25] (and hence the generating function formulation for arbitrary subgraphs) at the time of writing [33] and [32].

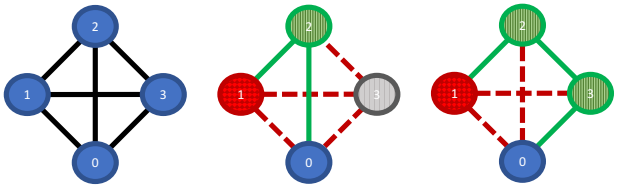


Figure 5.6: The 4-clique (left) with labelled vertex sites and focal vertex chosen to be vertex 0. Assuming that vertex 1 is attached to the GCC (diamond checked red vertex), then there are two types of non-direct walks back to the focal vertex. The 1-hop walk (center) requires that vertex 2 is not attached to the GCC (vertical green pattern). Bond occupation (solid edges) must occur through the path [1,2,0], which we term the success path. The state of vertex 3 is unspecified by the success path (grey checked pattern). However, all other paths, from any starting vertex, that do not cause intersection with the success path, must fail to attach vertex 0 to the GCC (red dashed edges). We term these the failure paths for the given success path under consideration. For the center success path, the failure paths are [1,3,0], [2,3,0] and [3,0]. The first two assume that vertex 3 is in state u_4, while the final path assumes vertex 3 was attached to the GCC prior to this.

distinct paths that can be made back to the focal vertex: 1-hop (center) and 2-hop (right) walks.

The second term in Eq 5.27 concerns the 1-hop walks in the clique. Consider (for instance) that vertex 1 is the source vertex. For this walk to occur vertex 1 must be attached to the GCC with probability $(1-u_4)$, but it has failed to attach the focal vertex directly with probability $1-T$. vertex 2 (for instance) must become attached through bond occupation from vertex 1 with probability u_4T, which then goes on to connect to the focal vertex through its direct edge with probability T. We then must ensure that all the remaining pieces in the clique that have not been assigned a probability must be dealt with, we cannot leave them unaccounted for. Both vertex 1 and vertex 2 must fail to exercise their alternative 1-hop walks back to the focal vertex. The probability of each of these walks failing is $1-u_4T^2$. However, it might happen that vertex 3 was also attached to the GCC, in which case, it must fail directly with probability $(1-u_4)(1-T)$. The factor of 6 accounts for the path multiplicity; each vertex has 2 1-hop walks back to the focal vertex. For instance, we depicted the success path in Fig 5.6 as [1,2,0], however, another valid 1-hop walk from 1 is [1,3,0].

The final term is much easier to rationalise. Consider again that vertex 1 is attached to the GCC, but that it has failed directly to connect to the focal vertex, Fig 5.6. The 2-hop walk [0,2,3,0] back to the focal vertex around the clique must fix both vertices 2 and 3 to be unattached and involve three bond occupation events. Further, both interior edges in the clique must not short-circuit the 2-hop walk into a 1-hop walk, so they too, in addition to vertex 1s direct edge, must be unoccupied. The other 2-hop walk starting from 1 is given by [1,3,2,0].

The method is abstracted to all clique sizes [32] where the magnitude of the error of the approximation is also discussed. A series of applications of this method were presented [33]; including multilayer models, arbitrary motifs and semi-directed networks.

The probabilities of not becoming part of the GCC through the 4-cycle sites, g_{4_3} and

g_{4_2} are given by

$$\begin{aligned} g_{4_3} = &[u_{4_2} + (1-u_{4_2})(1-T)]^2[u_{4_3} + (1-u_{4_3})(1-T)] \\ &- 2(1-u_{4_2})(1-T)^2 u_{4_2} u_{4_3} T^3 \\ &- 2(1-u_{4_2})(1-T) u_{4_3} T^2 (1-u_{4_2}T^2)(1-(1-u_{4_2})T) \\ &- 2(1-u_{4_3})(1-T) u_{4_2} T^2 (1-u_{4_2}T^2)(1-(1-u_{4_2})T) \end{aligned} \tag{5.28}$$

and

$$\begin{aligned} g_{4_2} = &[u_{4_3} + (1-u_{4_3})(1-T)]^2 \\ &- 2(1-T)^2(1-u_{4_3}) u_{4_2} u_{4_3} T^3 \\ &- 2(1-u_{4_2})(1-T)^2 u_{4_3}^2 T^3 \\ &- 2(1-u_{4_2}) u_{4_3} T^2 (1-u_{4_3}T^2)^2 (1-(1-u_{4_3})T) \\ &- 2(1-T)(1-u_{4_3}) u_{4_3} T^2 (1-u_{4_2}T^2)(1-(1-u_{4_2})T) \end{aligned} \tag{5.29}$$

These equations are understood as follows; firstly, we pick a unique vertex-site in the cycle as the focal vertex for the cycle under consideration. The first term in its g_τ equation is the product of probabilities that each direct-contact edge fails to connect it to the GCC. The leading term of g_{4_3} is cubic in $[u_\tau + (1-u_\tau)(1-T)]$ while g_{4_2} is quadratic. The remaining terms capture the probabilities that vertices use cycle-edges to connect the focal vertex to the GCC. An interesting observation is that this method essentially enumerates connected SIR trees of a given size in a clique motif. In reference to the percolation-SIR equivalence, there must also be an equivalence between the probability of all connected trees and all connected graphs on a clique; a relation we hope to investigate further.

5.5 Chapter summary

This chapter has discussed the application of the generating function formulation to GCM networks that are composed of non-clique subgraphs. This composed of applying Karrer and Newman's [25] generalised system of equations and finding the appropriate g_η expression, which is an NP hard problem. To find g_η, we applied the combinatorial counting algorithm which previously afforded the closed-form clique expression in chapter 3. We successfully found the percolation formula for chordless cycles in closed-form. It is unlikely, however, that a general closed-form expression could be found beyond the expressions presented in this chapter simply due to the complexity and uniqueness of each distinct motif. We believe there is more work to be done in characterising the effects of subtle changes in topology or edge arrangement for a family of subgraphs, with a given vertex count, that lie within the bounds of the chordless cycle and the clique. In the final section, we reviewed an alternative method to find g_η based on the SIR equivalence of finding connected infection trees within a motif. Whilst not exact, this method has proven to be very accurate for a range of networked phenomena including: semi-directed, modular, multilayer and networks with arbitrary motifs [32, 33].

CHAPTER SIX

DEGREE CORRELATIONS IN CLIQUE RANDOM NETWORKS

Correlations among the degrees of vertices in random graphs often occur when clustering is present. In this chapter we define a joint-degree correlation function for vertices in the giant component of clustered configuration model networks which are comprised of higher-order subgraphs. We use this model to investigate, in detail, the organisation among nearest-neighbour subgraphs for random graphs as a function of subgraph topology as well as clustering. We find an expression for the average joint degree of a neighbour in the giant component at the critical point for these networks. Finally, we introduce a novel edge-disjoint clique decomposition algorithm and investigate the correlations between the subgraphs of empirical networks. We compare our clique cover to other methods in the literature by examining the correlations between clique motifs of an empirical network. We find our method performs best when large cliques are present in the network.

6.1 Degree correlations in clique random networks

Consider an arbitrary set of edge topologies including ordinary edges, triangles, squares, 4-cliques, pentagons and so on, denoted by $\vec{\boldsymbol{\tau}} = \{\perp, \Delta, \square, \ldots, \gamma\}$, where γ is the topology of the final element. In the following, we reserve τ and ν as indices over elements of $\boldsymbol{\tau}$. We define the number of cycles that a vertex plays a role in for each topology $\tau \in \boldsymbol{\tau}$ by vector $\boldsymbol{k}_{\boldsymbol{\tau},l} = \{k_\perp, k_\Delta, \ldots, k_\gamma\}$ with $l = 0, 1$ representing the focal vertex and nearest-neighbour joint sequences, respectively. We reserve $k_{\nu,l} \in \boldsymbol{k}_{\boldsymbol{\tau},l}$ as an index for the number of cycles of topology ν around a given vertex in layer l; we drop the l label where obvious. The joint probability distribution for choosing this vertex at random is then denoted as $p_{\boldsymbol{k}_{\boldsymbol{\tau},l}}$. The number of edges that a given vertex has within each cycle is defined by m_τ; for instance a vertex contributes two edges to each triangle it connects to and hence $m_\Delta = 2$. We define n_{τ,ν,k_ν} to be the number of vertices with k_ν cycles of topology ν that we reach by following an edge of topology τ from the focal vertex to a nearest neighbour. There are $\dim(\vec{\tau}^2)$ of these expressions. Let a particular configuration of type ν following τ edges be $n_{\tau,\nu}$ such that

$$n_{\tau,\nu} = \{n_{\tau,\nu,1}, n_{\tau,\nu,2}, \ldots\} \tag{6.1}$$

Then, we define the set of all configurations of the neighbours following τ edges to be $n_\tau = \{n_{\tau,\perp}, n_{\tau,\Delta}, \ldots\}$. Finally, the set of all configurations is denoted by $n = \{n_\perp, n_\Delta, \ldots\}$. The number of vertices reached by following all of the τ edges is

$$N_\tau = \sum_{k_\tau=1} n_{\tau,\tau,k_\tau} = \sum_{k_\nu=0} n_{\tau,\nu,k_\nu} \qquad \tau \neq \nu \tag{6.2}$$

The total number of vertices 1-layer out from the focal vertex is the sum of all vertices reached by traversing each edge topology

$$N = \sum_{\tau \in \boldsymbol{\tau}} N_\tau \tag{6.3}$$

Let $P(n \mid N)$ be the probability that the nearest-neighbour configuration is given by set n and that the total number of vertices in the first layer is N. This is given by

$$P(n \mid N) = \prod_\tau \left(\prod_{\nu \neq \tau} \prod_{k_\nu = 0} \frac{N_\tau}{n_{\tau,\nu,k_\nu}!} q_{\tau,\nu,k_\nu}^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau = 1} \frac{N_\tau}{n_{\tau,\tau,k_\tau}!} q_{\tau,\tau,k_\tau}^{n_{\tau,\tau,k_\tau}} \tag{6.4}$$

where $q_{\tau,\nu,k}$ is the probability of traversing an edge of topology τ to a vertex with k_ν independent cycles of topology ν. We also have the understanding that each term of the product over $\nu \neq \tau$ has its own index k_ν starting from zero; we have pulled out τ from this expression since, by definition, there must be at least one τ-edge present to follow it to a nearest neighbour vertex and so the index starts at 1. The probability $P(\mathrm{GC} \mid n)$ that the component is the GCC for a particular configuration n is given by

$$P(\mathrm{GC} \mid n, N) = 1 - \prod_\tau \left(\prod_{\nu \neq \tau} \prod_{k_\nu = 0} \left[u_\nu^{m_\nu k_\nu} \right]^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau = 1} \left[u_\tau^{m_\tau (k_\tau - 1)} \right]^{n_{\tau,\tau,k_\tau}} \tag{6.5}$$

where we have introduced u_τ as the probability that a vertex at the end of a randomly chosen edge of topology τ fails to connect to the GC. The probability that the configuration is n, that the component is the GCC given that there are N nearest-neighbours is found

from Bayes' theorem as

$$P(n,\mathrm{GC} \mid N) = P(\mathrm{GC} \mid n,N)P(n \mid N) \tag{6.6}$$

Let $P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0})$ be the probability of there being N vertices in the 1st layer given that the joint degree of the focal vertex is $\boldsymbol{k}_{\boldsymbol{\tau},0}$ and that the component is the GC. We can use this to find the probability $P(n,\mathrm{GC} \mid \boldsymbol{k}_{\boldsymbol{\tau},0})$ that the nearest-neighbour configuration is n given the joint degree of a vertex in the GCC is $\boldsymbol{k}_{\boldsymbol{\tau},0}$ as

$$P(n,\mathrm{GC} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \sum_{N} P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0})P(n,\mathrm{GC} \mid N) \tag{6.7}$$

where the summation is over all combinations of N_τ such that

$$\sum_{N} = \sum_{N_\perp}\sum_{N_\Delta}\cdots \tag{6.8}$$

We find

$$\begin{aligned} P(n,\mathrm{GC} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \sum_{N} P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_{\tau} \left(\prod_{\nu\neq\tau} \prod_{k_\nu=0} \frac{N_\tau}{n_{\tau,\nu,k_\nu}!} q_{\tau,\nu,k_\nu}^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau=1} \frac{N_\tau}{n_{\tau,\tau,k_\tau}!} q_{\tau,\tau,k_\tau}^{n_{\tau,\tau,k_\tau}} \\ \times \left[1 - \prod_{\eta} \left(\prod_{\varphi\neq\eta} \prod_{k_\nu=0} \left[u_\varphi^{m_\varphi k_\nu} \right]^{n_{\eta,\varphi,k_\nu}} \right) \prod_{k_\tau=1} [u_\eta^{m_\eta(k_\tau-1)}]^{n_{\eta,\eta,k_\tau}} \right] \end{aligned} \tag{6.9}$$

for $\tau,\nu,\eta,\varphi \in \boldsymbol{\tau}$. We now generate this probability by summing over all permissible configurations of the nearest-neighbour joint degrees to obtain

$$\tilde{F}_{\mathrm{GC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \sum_{n} P(n,\mathrm{GC} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_{\tau} \left(\prod_{\nu\neq\tau} \prod_{k_\nu=0} X_{\tau,\nu,k_\nu}^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau=1} X_{\tau,\tau,k_\tau}^{n_{\tau,\tau,k_\tau}} \tag{6.10}$$

where

$$\sum_{n} = \sum_{n_{2,\perp}} \sum_{n_{\perp,\Delta}} \cdots \sum_{n_{\Delta,\perp}} \sum_{n_{\Delta,\Delta}} \cdots \tag{6.11}$$

We simplify the expression by substituting Eq 6.9, swapping the order of the summations and collecting terms in like powers to obtain

$$\begin{aligned} \tilde{F}_{\mathrm{GC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \sum_{n}\sum_{N} P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_{\tau} \left(\prod_{\nu\neq\tau} \prod_{k_\nu=0} \frac{N_\tau}{n_{\tau,\nu,k_\nu}!} q_{\tau,\nu,k_\nu}^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau=1} \frac{N_\tau}{n_{\tau,\tau,k_\tau}!} q_{\tau,\tau,k_\tau}^{n_{\tau,\tau,k_\tau}} \\ \times \left[1 - \prod_{\eta} \left(\prod_{\varphi\neq\eta} \prod_{k_\nu=0} [u_\varphi^{m_\varphi k_\nu}]^{n_{\eta,\varphi,k_\nu}} \right) \prod_{k_\tau=1} [u_\eta^{m_\eta(k_\tau-1)}]^{n_{\eta,\eta,k_\tau}} \right] \\ \times \prod_{\tau} \left(\prod_{\nu\neq\tau} \prod_{k_\nu=0} X_{\tau,\nu,k_\nu}^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau=1} X_{\tau,\tau,k_\tau}^{n_{\tau,\tau,k_\tau}} \end{aligned} \tag{6.12}$$

to find

$$\tilde{F}_{\mathrm{GC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \sum_n \sum_N P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_\tau \left(\prod_{\nu \neq \tau} \prod_{k_\nu=0} \frac{N_\tau}{n_{\tau,\nu,k_\nu}!} (q_{\tau,\nu,k_\nu} X_{\tau,\nu,k_\nu})^{n_{\tau,\nu,k_\nu}} \right) \times \prod_{k_\tau=1} \frac{N_\tau}{n_{\tau,\tau,k_\tau}!} (q_{\tau,\tau,k_\tau} X_{\tau,\tau,k_\tau})^{n_{\tau,\tau,k_\tau}} \times \left[1 - \prod_\eta \left(\prod_{\varphi \neq \eta} \prod_{k_\nu=0} \left[u_\varphi^{m_\varphi k_\nu} \right]^{n_{\eta,\varphi,k_\nu}} \right) \prod_{k_\tau=1} [u_\eta^{m_\eta (k_\tau - 1)}]^{n_{\eta,\eta,k_\tau}} \right] \tag{6.13}$$

The multinomial theorem can now be applied to each of the terms in the product to obtain

$$\begin{aligned} \tilde{F}_{\mathrm{GC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = {} & \sum_N P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_\tau \left[\left(\prod_{\nu \neq \tau} \sum_{k_\nu=0} q_{\tau,\nu,k_\nu} X_{\tau,\nu,k_\nu} \right) \sum_{k_\tau=1} q_{\tau,\tau,k_\tau} X_{\tau,\tau,k_\tau} \right]^{N_\tau} \\ & - \sum_N P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_\tau \\ & \times \left[\left(\prod_{\nu \neq \tau} \sum_{k_\nu=0} q_{\tau,\nu,k_\nu} u_\nu^{m_\nu k_\nu} X_{\tau,\nu,k_\nu} \right) \sum_{k_\tau=1} q_{\tau,\tau,k_\tau} u_\tau^{m_\tau (k_\tau - 1)} X_{\tau,\tau,k_\tau} \right]^{N_\tau} \end{aligned} \tag{6.14}$$

The probability that an edge of topology τ can be followed to reach a vertex with k_ν cycles of topology ν is given by q_{τ,ν,k_ν}. The probability that an edge of topology τ can be traversed to reach a vertex with k_ν cycles of topology ν for all $\nu \in \boldsymbol{\tau}$ is the joint excess degree distribution, $q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},l}}$. This can be constructed from the separable distributions such that

$$q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},l}} = \prod_\nu q_{\tau,\nu,k_{\nu,l}} \tag{6.15}$$

With this we can write

$$\begin{aligned} \tilde{F}_{\mathrm{GC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = {} & \sum_N P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_\tau \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{N_\tau} \\ & - \sum_N P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_\tau \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} u_\nu^{m_\nu k_\nu} u_\tau^{m_\tau (k_\tau - 1)} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{N_\tau} \end{aligned} \tag{6.16}$$

The probability that there are N nearest-neighbour vertices given the joint degree of the focal vertex is $\boldsymbol{k}_{\boldsymbol{\tau},0}$ is simply a particular term from the $G_0(\boldsymbol{Z})$ generating function. Inserting this definition into our expression we arrive at the generating function that describes the

distribution of nearest-neighbours given a particular joint degree of the focal vertex as

$$\hat{F}_{\mathrm{GC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} \prod_{\tau} \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{m_\tau k_{\tau,0}} - p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} \prod_{\tau} \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} u_\nu^{m_\nu k_\nu} u_\tau^{m_\tau(k_\tau-1)} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{m_\tau k_{\tau,0}} \tag{6.17}$$

The expectation number of the number of nearest-neighbours with a given joint degree is found from the expectation value of $\hat{F}_{\mathrm{GC}}(\boldsymbol{X} = Z \mid \boldsymbol{k}_{\boldsymbol{\tau},0})$. We then find

$$\hat{F}'_{\mathrm{GC}} = \sum_{\tau \in \boldsymbol{\tau}} m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} \left(1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1} \prod_{\nu \in \boldsymbol{\tau} \setminus \tau} u_\nu^{m_\nu(k_{\nu,0}+k_{\nu,1})} \right) \tag{6.18}$$

where the derivative is evaluated at $Z_{\boldsymbol{k}_{\tau,1}} = 1$ (see appendix C for a complete derivation using the tree-triangle model). The bracket is one minus the probability that the none of the edges to the second layer lead to the GC; whilst the prefactor describes the probability of following $k_{\tau,0}$ τ-cycles, each of which has m_τ edges to follow to reach a vertex whose joint degree is given by $q_{\tau,\boldsymbol{k}_{\tau,1}}$.

In a similar way, we can find the generating function $F_{\mathrm{GC}}(\boldsymbol{X})$ for the probability distribution that a randomly chosen vertex has a nearest neighbour configuration given by n and belongs to the GCC as

$$F_{\mathrm{GC}}(\boldsymbol{X}) = \sum_{\boldsymbol{k}_{\boldsymbol{\tau},0}} \hat{F}_{\mathrm{GC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \tag{6.19}$$

$$= \sum_{\boldsymbol{k}_{\boldsymbol{\tau},0}} p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} \prod_{\tau} \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{m_\tau k_{\tau,0}} - \sum_{\boldsymbol{k}_{\boldsymbol{\tau},0}} p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} \prod_{\tau} \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} u_\nu^{m_\nu k_\nu} u_\tau^{m_\tau(k_\tau-1)} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{m_\tau k_{\tau,0}} \tag{6.20}$$

which is simply $G_0(\boldsymbol{Z})$. The expectation number for the of nearest-neighbours from a random focal vertex in the GCC is given by

$$F'_{\mathrm{GC}} = \sum_{\tau \in \boldsymbol{\tau}} m_\tau \langle k_\tau \rangle [1 - u_\tau^{m_\tau \omega_\tau}] \tag{6.21}$$

where ω_τ represents the number of vertices in the cycle. We can use the quotient of these expectation values to define a symmetric joint-probability distribution $P_{\mathrm{GC}}(\boldsymbol{k}_{\boldsymbol{\tau},0}, \boldsymbol{k}_{\boldsymbol{\tau},1}) = \hat{F}'_{\mathrm{GC}}/F'_{\mathrm{GC}}$ that two nearest-neighbours in the GCC have joint degrees $\boldsymbol{k}_{\boldsymbol{\tau},0}$ and $\boldsymbol{k}_{\boldsymbol{\tau},1}$ as

$$P_{\mathrm{GC}}(\boldsymbol{k}_{\boldsymbol{\tau},0}, \boldsymbol{k}_{\boldsymbol{\tau},1}) = \sum_{\tau \in \boldsymbol{\tau}} m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} \left(1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1} \times \prod_{\nu \in \boldsymbol{\tau} \setminus \tau} u_\nu^{m_\nu(k_{\nu,0}+k_{\nu,1})} \right) \Big/ \sum_{\tau \in \boldsymbol{\tau}} m_\tau \langle k_\tau \rangle [1 - u_\tau^{m_\tau \omega_\tau}] \tag{6.22}$$

where $P_{\mathrm{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},0},\boldsymbol{k}_{\boldsymbol{\tau},1}) = P_{\mathrm{GCC}}(k_{\perp,0},\ldots,k_{\gamma,0},k_{\perp,1},\ldots,k_{\gamma,1})$. This equation is a central result and can be used to compute many interesting properties of the correlation structure within configuration model networks. At any time, we can compress the information contained within $P_{\mathrm{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},0},\boldsymbol{k}_{\boldsymbol{\tau},1})$ to find $P_{\mathrm{GCC}}(k_0,k_1)$ which is the probability that a focal vertex with overall degree k_0 attaches to a neighbour whose overall degree is k_1.

$$P_{\mathrm{GCC}}^{\mathrm{overall}}(k_0,k_1) = \sum_{\tau}\sum_{k_\tau} P_{\mathrm{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},0},\boldsymbol{k}_{\boldsymbol{\tau},1})\delta_{k_0,k_0^{\mathrm{overall}}}\delta_{k_1,k_0^{\mathrm{overall}}} \tag{6.23}$$

where $k_0^{\mathrm{overall}} = \sum_\tau \sum_{k_{\tau,0}} m_\tau k_{\tau,0}$ and $k_1^{\mathrm{overall}} = \sum_\tau \sum_{k_{\tau,1}} m_\tau k_{\tau,1}$ are the overall degrees of the focal and neighbour vertices. However, this degree lumping procedure overlooks the fine structure among the correlations as many joint degrees can contribute to a given overall degree. Indeed it is precisely this structure which acts as a fingerprint of a network ensemble.

Let us introduce the conditional probability

$$P_{\mathrm{GC}}(k_{\perp,1},\ldots,k_{\gamma,1} \mid k_{\perp,0},\ldots,k_{\gamma,0}) = P_{\mathrm{GC}}(\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \tag{6.24}$$

that the nearest neighbour has joint degree $\boldsymbol{k}_{\boldsymbol{\tau},1}$ given that the focal vertex has joint degree $\boldsymbol{k}_{\boldsymbol{\tau},0}$ in the GC. Applying Bayes' theorem to our discrete multivariate joint probability we have

$$P_{\mathrm{GC}}(\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \frac{P_{\mathrm{GC}}(k_{\perp,0},\ldots,k_{\gamma,0} \mid k_{\perp,1},\ldots,k_{\gamma,1})P_{\mathrm{GC}}(k_{\perp,1},\ldots,k_{\gamma,1})}{\sum_{k_{\perp,1},\ldots,k_{\gamma,1}} P_{\mathrm{GC}}(k_{\perp,0},\ldots,k_{\gamma,0} \mid k_{\perp,1},\ldots,k_{\gamma,1})P_{\mathrm{GC}}(k_{\perp,1},\ldots,k_{\gamma,1})} \tag{6.25}$$

Which simplifies to

$$P_{\mathrm{GC}}(\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \frac{P_{\mathrm{GC}}(\boldsymbol{k}_{\boldsymbol{\tau},0},\boldsymbol{k}_{\boldsymbol{\tau},1})}{\sum_{\boldsymbol{k}_{\boldsymbol{\tau},1}} P_{\mathrm{GC}}(\boldsymbol{k}_{\boldsymbol{\tau},0},\boldsymbol{k}_{\boldsymbol{\tau},1})} \tag{6.26}$$

Inserting Eq 6.22 we find

$$P_{\mathrm{GC}}(\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \frac{\sum_{\tau\in\boldsymbol{\tau}} m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} \left(1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1} \prod_{\nu\in\boldsymbol{\tau}\setminus\tau} u_\nu^{m_\nu(k_{\nu,0}+k_{\nu,1})}\right)}{\sum_{\tau\in\boldsymbol{\tau}}\sum_{k_{\tau,1}} m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} \left(1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1} \prod_{\nu\in\boldsymbol{\tau}\setminus\tau} u_\nu^{m_\nu(k_{\nu,0}+k_{\nu,1})}\right)} \tag{6.27}$$

We can use $P_{\mathrm{GC}}(\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0})$ to find multivariate conditional expectation values for a given focal vertex joint degree, generalising [57] for the GCM. The expectation value for vector $\boldsymbol{X}$ given vector $\boldsymbol{Y}$ is a vector $E[\boldsymbol{X} \mid \boldsymbol{Y}] = (E[X_1 \mid Y],\ldots,E[X_n \mid \boldsymbol{Y}])^{\mathrm{T}}$ whose elements are the expected values of each of the variables defined as

$$E[X_i \mid \boldsymbol{Y} = \boldsymbol{y}] = \sum_{x_1,\ldots,x_n} x_i P_{\mathrm{GC}}(x_1,\ldots,x_n \mid \boldsymbol{Y} = \boldsymbol{y}) \tag{6.28}$$

For instance, the average joint degree of a neighbour to a focal vertex whose joint degree is $\boldsymbol{k}_{\tau,0}$ is the vector $(E[k_{\perp,1} \mid \boldsymbol{k}_{\tau,0}],\ldots,E[k_{\gamma,1} \mid \boldsymbol{k}_{\tau,0}])^T$ whose elements are

$$E[k_{\tau,1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}] = \sum_{\boldsymbol{k}_{\boldsymbol{\tau},1}} k_{\tau,1} P(\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \tag{6.29}$$

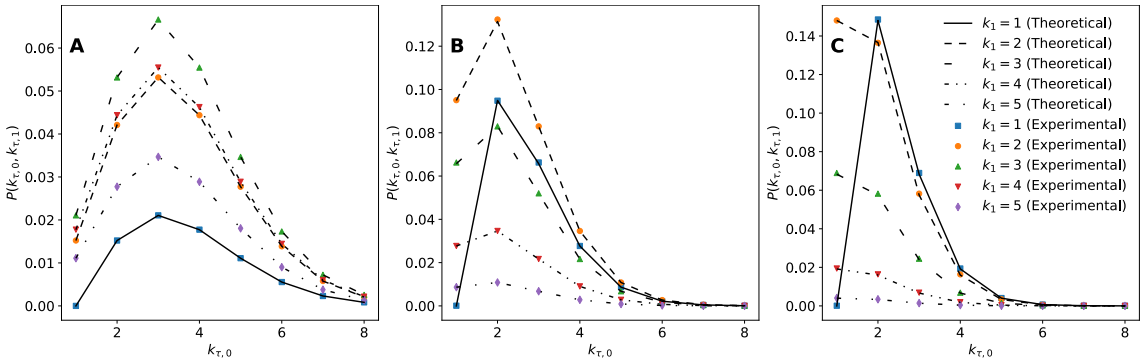


Figure 6.1: The probability $P(k_{\tau,0},k_{\tau,1})$ for Erdős-Renyi random graphs comprising of a single motif topology, 2-cliques (A), 3-cliques (B) and 4-cliques (C), respectively, as a function of $k_{\tau,0}$ for several $k_{\tau,1}$. The overall mean degree is fixed at $\langle k\rangle = 2.5$ for networks with $N = 60000$ vertices. Scatter points are the average of 100 repetitions of Monte Carlo simulation while the lines are the theoretical predictions from Eq 6.30. The legend is the same for each plot.

We examine this expression in Appendix C for the tree-triangle model.

In this chapter we have introduced a theoretical model, based on generating functions, to investigate the NNDC in the GCC of random clustered graphs, constructed according to the GCM, comprising of higher-order clusters. We now examine a series of pertinent examples of this model.

6.1.1 Single topology

In the special case that the network consists of a single homogeneous subgraph (a homogeneous subgraph is one where all vertices are degree-regular), then $P_{\mathrm{GC}}(k_{\tau,0},k_{\tau,1})$ from Eq 6.22 is given by

$$P_{\mathrm{GC}}(k_{\tau,0},k_{\tau,1}) = \frac{(1-u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1})}{1-u_\tau^{m_\tau\omega_\tau}} q_{\tau,k_{\tau,0}} q_{\tau,k_{\tau,1}} \tag{6.30}$$

and similarly from Eq 6.27 we have the related conditional probability

$$P_{\mathrm{GC}}(k_{\tau,1} \mid k_{\tau,0}) = \frac{(1-u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1})}{1-u_\tau^{m_\tau k_{\tau,0}}} q_{\tau,k_{\tau,1}} \tag{6.31}$$

which reproduces the results of [2, 65] for the nearest-neighbour distributions on the GCC of tree-like networks when $\tau = \perp$. We examine the NNDCs for single-topology networks with Poisson distribution participation in motifs with fixed mean degree $\langle k\rangle = 2.5$ in Fig 6.1. The networks are composed of discrete clique topologies; specifically 2, 3 and 4-cliques in Fig 6.1 A, B and C, respectively. The markers are the averaged results of Monte Carlo simulation while the lines are the theoretical predictions of Eq 6.30; both are in excellent agreement. In each case, $P_{\mathrm{GC}}(k_{\tau,0},k_{\tau,1})$ is plotted as a function of increasing $k_{\tau,0}$ for several $k_{\tau,1}$ values. We note that for each clique size $P_{\mathrm{GC}}(1,1) = 0$; since, this combination cannot exist in the GC. For networks comprised of a single topology, the

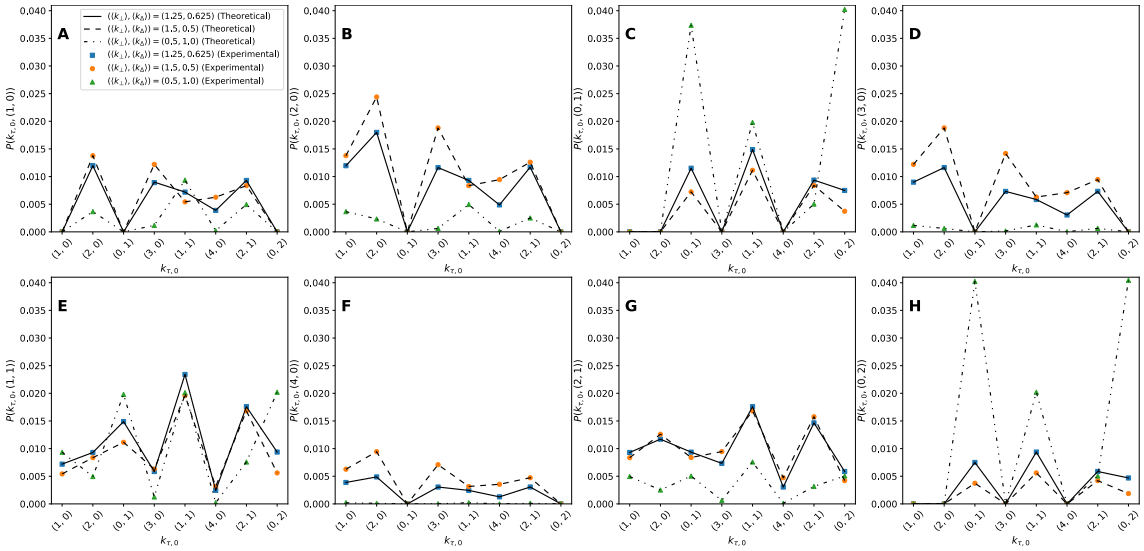


Figure 6.2: The probability $P_{\mathrm{GCC}}(s_0,t_0,s_1,t_1)$ for Poisson random graphs comprising of mixed 2-clique and 3-clique topologies for three different clustering regimes. In each plot, the joint degrees of the focal vertex up to overall degree $k=4$ are plotted on the horizontal axis for a given (s_1,t_1) neighbour. Scatter points are the average of 250 repetitions of Monte Carlo simulation on networks with 2×10^5 vertices; whilst lines are the analytical results of Eq 6.22. The legend is the same as tile (A) for all plots.

average degree of a neighbour can be found from Eq 6.29 as

$$E[k_{\tau,1} \mid k_{\tau,0}] = \frac{\sum\limits_{k_{\tau,1}} k_{\tau,1} q_{\tau,k_{\tau,1}} (1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1})}{1 - u_\tau^{m_\tau k_{\tau,0}}} \tag{6.32}$$

which is in agreement with [41] for tree-like topologies.

6.1.2 Tree-triangle model

We now examine how clustering influences the degree correlations in the GCC of the mixed topology tree-triangle model. The theoretical details of this model are derived in Appendix 8.3. Fixing the first moment of the model to $\langle k \rangle = 2.5$ the limiting cases of $\langle k_\perp \rangle = 0$ and $\langle k_\Delta \rangle = 0$ are presented in Fig 6.1 and we now examine i) an even neighbour distribution by setting $\langle k_\perp \rangle = 1.25$ and $\langle k_\Delta \rangle = 0.625$; ii) a weakly clustered regime with $\langle k_\perp \rangle = 1.5$ and $\langle k_\Delta \rangle = 0.5$ and finally iii) a strong clustering regime with $\langle k_\perp \rangle = 0.5$ and $\langle k_\Delta \rangle = 1.0$ in Fig 6.2. The joint degree of the horizontal axis is ordered by increasing overall degree. When a given overall degree can be formed in multiple ways, such as $k=2$ from $(2,0)$ or $(0,1)$, the degenerate cases are ordered by increasing local clustering coefficient. Each tile in Fig 6.2 A-H plots a given neighbour joint degree (as a function of the focal vertex joint degree) for the three clustering regimes. We observe some encouraging results from these plots: firstly, as with the results of experiments with single-topology networks (Fig 6.1), the probabilities $P_{\mathrm{GCC}}(1,0,1,0)$ and $P_{\mathrm{GCC}}(0,1,0,1)$ are both zero for the vertices in the GCC (see Fig 6.2 A). We also notice that $P_{\mathrm{GCC}}(s_0,t_0,s_1,t_1)$ takes zero values for impossible combinations, such as neighbours whose edges are of a single, yet opposite, topology to one another.

Further, the probabilities are symmetric such that $P_{\mathrm{GCC}}(k_{\boldsymbol{\tau},0}, k_{\boldsymbol{\tau},1}) = P_{\mathrm{GCC}}(k_{\boldsymbol{\tau},1}, k_{\boldsymbol{\tau},0})$ which is an expected result for undirected random graphs. Among the non-zero combinations we observe that some peaks, particularly among focal vertices with non-zero degrees in both topologies, are aligned across all series; for example $P_{\mathrm{GCC}}(1,1,1,1)$ in E. Conversely, other peaks such as $P_{\mathrm{GCC}}(2,0,2,1)$ in G peak in the weak and even regimes, yet trough in the strong clustered regime.

We also observe, across all tiles in Fig 6.2 that the correlations among the weak (blue squares) and even-neighbour (orange circles) regimes are generally of higher magnitude across all focal vertices than the strongly clustered regime (green triangles). In other words, the networks with strong clustering exhibit NNDC that have smaller magnitudes with the exception of tiles C and H, which consider neighbouring vertices that only have triangle motifs.

In tile F we notice that vertices with a high tree-like degree do not tend to connect with neighbours with triangles, especially in the strong clustering regime.

Collectively, these results give insight into how the network is held together at the microscopic level and how the presence of clustering alters this structure. This could prove useful for creating synthetic networks or for a better understanding of network resilience under targeted attack.

6.1.3 The effect of clique size on NNDC

In this section, we examine the effect of increasing the clique size on the NNDC of mixed topology GCM networks. To achieve this, we extend the calculations performed in appendix 8.3 from the 2- and 3-clique model to a binary model composed of 2- and m-cliques, whose topology we denote by σ. For this model, the NNDC for a focal vertex with s_0 ordinary edges and c_0 edge-disjoint m-cliques in the GCC of a GCM network can be obtained from

$$\begin{aligned} P_{\mathrm{GCC}}(s_0,c_0,s',c') = &\Big[p_{s_0c_0} s_0 q_{\perp,(s',c')} \left[1 - u_\perp^{s_0+s'-2} u_\sigma^{m_\sigma(c_0+c')}\right] \\ &+ m_\sigma c_0 p_{s_0c_0} q_{\sigma,(s',c')} \left[1 - u_\perp^{s_0+s'} u_\sigma^{m_\sigma(c_0+c'-1)-1}\right] \Big] \Big/ \Big[\langle s\rangle (1-u_\perp^2) \\ &+ m_\sigma \langle c\rangle (1-u_\sigma^{\omega_\sigma}) \Big] \end{aligned} \tag{6.33}$$

The results of this expression are shown in Fig 6.3, where the overall neighbour degree is plotted against the overall degree of the focal vertex for several increasing clique sizes. The scatter points are the results of Monte Carlo simulation of networks with 100000 vertices, whilst the plotted lines are the theoretical results of the model; both show excellent agreement with one another. The networks are constructed according to the GCM algorithm before the GCC is selected from the possibly disconnected graph. The motifs counts at each vertex are drawn from Poisson distributions with averages chosen such that the first moment of the distribution of overall degrees is fixed at $\langle k\rangle = 6$ across all experiments whilst the average 2-clique count is held fixed at $\langle k_\perp\rangle = 1.25$ and the average clique count $\langle k_\sigma\rangle$ is the solution of $\langle k\rangle = \langle k_\perp\rangle + m_\sigma\langle k_\sigma\rangle$. From Fig 6.3 we observe that the average neighbour degree of networks with larger cliques increases. For cliques larger than 2-cliques, oscillations in the average neighbour degree appear at low focal vertex degree.

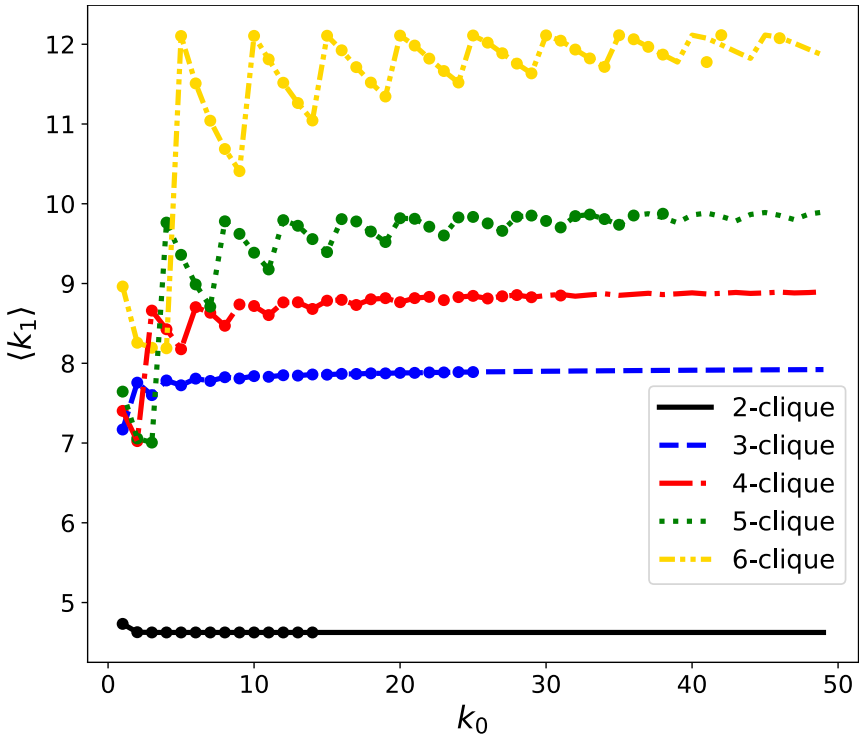


Figure 6.3: The average overall degree of a neighbour for increasing focal vertex degree for binary-topology networks comprising 2-cliques and higher-order cliques. Scatter points are the average of 1000 repetitions of Monte Carlo simulation whilst the plotted lines are the result Eq 6.33, collected by overall degree according to Eq 6.23. The networks are created from the GCM algorithm with Poisson marginal distributions of each motif topology and overall average degree fixed at $\langle k \rangle = 6$ with $\langle k_\perp \rangle = 1.25$ across all experiments.

The amplitude of the oscillations increases with clique size. In each case, the oscillations dampen to a fixed value in the limit of large focal vertex degree.

6.1.4 Emergence of correlations

At criticality, as the GCC emerges, we have that $u_\tau \to 1$; the probability of not belonging to the GCC is near unity. In this case, the multivariate limit of Eq 6.22 does not exist. However, in the case that the network is composed of cliques of various sizes which are each independently Poisson distributed at each vertex such that

$$p_{\boldsymbol{k}_{\boldsymbol{\tau},l}} = q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},l}} = \prod_{\tau \in \boldsymbol{\tau}} e^{-\langle k_\tau \rangle} \frac{\langle k_\tau \rangle^{k_{\tau,l}}}{k_{\tau,l}!} \qquad \forall \tau \in \boldsymbol{\tau} \tag{6.34}$$

we have that $u_\tau = u^{m_\tau}, \forall \tau$ [25]. In this instance Eq 6.22 is a univariate distribution and we can use L'Hôpital's rule to determine the expected limit to be

$$\lim_{u \to 1} P_{\mathrm{GCC}}(k_{\tau,0}, k_{\tau,1}) = \frac{\sum_\tau m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} \Lambda_\tau q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}}}{\sum_\tau m_\tau^2 \omega_\tau \langle k_\tau \rangle} \tag{6.35}$$

where

$$\Lambda_\tau = m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1+\sum_{\nu\neq\tau} m_\nu(k_{\nu,0}+k_{\nu,1}) \tag{6.36}$$

The critical point can be found by linearising $u_\tau = G_{1,\tau}(\boldsymbol{u}_{\boldsymbol{\tau}}^{m_\tau})$ in a small perturbation ε around $u_\tau = 1-\varepsilon_\tau$ [32]. To leading order in the small parameter ε_τ we have $\boldsymbol{\varepsilon} = \boldsymbol{A\varepsilon}$ with $\boldsymbol{\varepsilon} = [\varepsilon_\perp, \varepsilon_\Delta, \ldots]^{\mathrm{T}}$. The GCC forms at the point when the determinant $\det|A-I|$ vanishes, where $A = [\partial \mathrm{G}/\partial u_\tau]$, $\mathrm{G} = [G_{1,\tau}, G_{1,\Delta}, \ldots, G_{1,\gamma}]$ and identity matrix I. With mixed topology networks a GCC can form in many different ways. For instance, the GCC of a random graph model with two topologies can form by three distinct mechanisms: a GCC can emerge solely in either of the topologies or global connectivity can occur through a mixture of the binary topologies.

As we approach the critical point from below, we introduce a characteristic scale κ_τ [64] associated to the joint degrees of the focal vertex and a neighbour given by $u_\tau = e^{-1/\kappa_\tau}$. Inserting this expression into Eq 6.22 for finite κ_τ in each topology, the correlations fall exponentially with increasing κ_τ and hence $P_{\mathrm{GCC}}(k_{\tau,0}, k_{\tau,1})$ tends to the uncorrelated value of

$$\sum_{\tau\in\boldsymbol{\tau}} m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} / \sum_{\tau\in\boldsymbol{\tau}} m_\tau \langle k_\tau\rangle \tag{6.37}$$

Therefore, when the joint degree exceeds the characteristic scale, the GCC is uncorrelated. It is clear that as u_τ approaches unity the scale diverges $\kappa_\tau \to \infty$ and hence, the GCC always exhibits degree correlations. In addition, approaching the critical point, the average joint degree (Eq 6.32) falls exponentially with increasing degree along each topology for fixed κ_τ.

$$\mathscr{E}[k_{\tau,1} \mid k_{\tau,0}] = \frac{\sum_{k_{\tau,1}} k_{\tau,1} q_{\tau,k_{\tau,1}} (1-e^{-\phi})}{1-e^{-m_\tau k_{\tau,0}/\kappa_\tau}} \tag{6.38}$$

where $\phi = m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1/\kappa_\tau$. Thus, the correlations which are present at the critical point are negative in nature. It might happen, however, given the number of ways that the GCC of a mixed motif random graph model can emerge, that the characteristic scales of all topologies don't diverge at the critical point. For instance, consider a doubly Poisson distributed tree-triangle model with a critical average tree degree, but a sub-critical average triangle degree. A GCC will form among the tree edges, but the probability of those vertices involved only in triangles, $(0,t)$ for $t = 1,2,3,\ldots$, connecting to this GCC is small; since, their connection requires them to connect to mixed-topology vertices, which in turn connect to the GCC. Thus, we might find that the negative degree correlation structure among the triangles has not yet formed despite there being a non-zero density of triangles in the GCC.

6.1.5 Empirical networks

We now examine the correlation properties of the GCC of the ensemble representation of empirical networks using our joint degree model. Random graphs are elements of an ensemble $\mathscr{G}$ of graphs with V vertices and E edges; each member occurring with probability $P(G)$ [2]. The average value of a property of graph G, $Z(G)$, (such as its degree distribution or average degree) can be averaged over the entire ensemble

$$\langle Z\rangle = \sum_{G\in\mathscr{G}} Z(G)P(G) \tag{6.39}$$

The generating function formulation describes the properties of the ensemble. Empirical networks G are particular realisations of members of $\mathscr{G}$. The properties of a particular realisation are given by

$$P(Z) = \sum_{G \in \mathscr{G}} \delta(Z - Z(G)) P(G) \tag{6.40}$$

If $P(Z)$ is well represented by the ensemble average then the generating function formulation can be used to describe the properties of G. To study the NNDC in the GCC of g using generating functions, we must represent the largest component of an empirical network by a joint degree sequence of subgraphs. Whilst the choice of subgraphs is arbitrary [33], we only include cliques in the topology representation due to the vast literature on clique finding algorithms and the simplicity of calculating their properties. The clique decomposition of the GCC of g whose cliques have order less than or equal to ω can be performed in many different ways; and the resulting joint degree sequence can exhibit significantly different properties in terms of the number of subgraphs present their clustering, and other properties. Given that the method to create the joint degree distribution is not unique, and that the ensemble properties of each particular decomposition are often dissimilar, we now examine three clique decompositions and compare their properties.

The trivial decomposition is to simply cover g with 2-cliques; we refer to this as the single-edge-decomposition (SED). The degree sequence can then be used to create realisations using the ordinary configuration model. Another simple cover is the minimal cover of maximal cliques. However, it is very likely that the edges of the cliques will not be disjoint, i.e. a single edge will be a member of more than one clique. Whilst this could be an accurate representation of a vertex's local environment, the construction process for random graphs using the GCM will not work. Thus, we must impose that the cover is edge-disjoint.

One proposed method of clique decomposition is defined heuristically as follows [6]: we obtain the set C of all maximal cliques from the network; each maximal n-clique $c_i \in C$, $n \in \{1, \ldots, \omega\}$ is scored according to the fraction of edges it shares with other members of C. The largest clique within the set of lowest score cliques are included in the representation and C is recalculated. The process is repeated until the edges of the substrate network are expended. Such a covering is known as a edge-disjoint edge clique cover (EECC), see Fig 6.4 for details. We propose a novel clique cover as follows: the set C of all cliques present in the network (including those induced from subgraphs of larger cliques) is obtained from the empirical network. The set is ordered such that the largest cliques have the highest precedence. The subset of cliques within C that have equal size $\forall n \in \{1, \ldots, \omega\}$ are then randomised; thus the cover is a Monte Carlo method. The largest cliques are drawn from C and placed on the network if their edges do not overlap other with cliques that have already been placed in the network. The list is iterated until all edges belong to an independent clique. This method draws non-maximal joint degree sequences; however, higher-order cliques are preferentially preserved, we describe it as an edge disjoint motif preserving edge clique cover (MPCC), see Fig 6.5. In the particular case that the set of maximal cliques are edge disjoint, the distribution obtained from both the EECC and MPCC motif decomposition algorithms are in agreement with one another. It should be mentioned that both covers are not unique when two cliques of a given size can be chosen. Within the MPCC, we resolve these degeneracies by retaining the cliques associated with higher degree vertices. In our implementation of the EECC, we choose cliques from the set of degenerate cliques at random.

Once a suitable cover has been formed for the network, its joint-degree sequence can

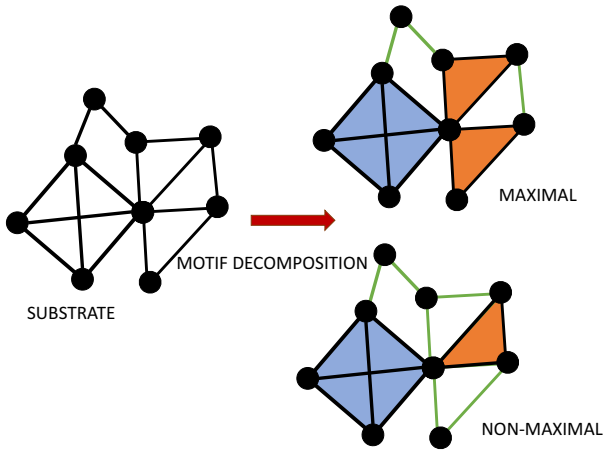


Figure 6.4: The clique decomposition of a substrate network (left) can be performed in multiple ways. Two examples are shown (right). The shaded faces are higher-order cliques whilst the green edges are 2-cliques. The clustering of the resulting joint degree distributions (and their random graph ensembles) are significantly altered depending on how the decomposition is performed. The maximal representation has 6 cliques in total whilst the non-maximal representation has 8 cliques. When only maximal representations are extracted the decomposition is a EECC.

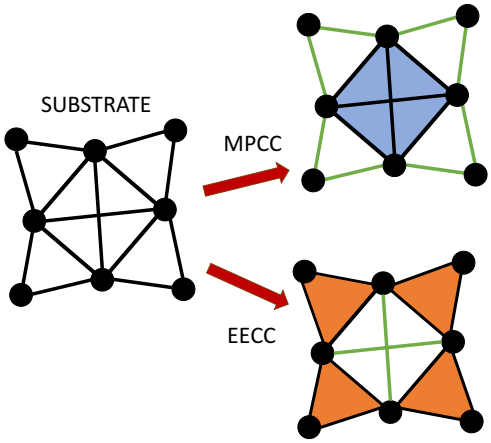


Figure 6.5: The results of the two clique decomposition algorithms (MPCC) and (EECC) for a particular substrate graph. The MPCC favours the formation of large cycles, leading to 9 cliques (a single 4-clique and 8 2-cliques) whilst the EECC leads to 6 cliques (4 3-cliques and 2 2-cliques). The joint degree sequence obtained from the MPCC network creates a non-maximal random ensemble of GCM networks.

be extracted. This sequence is then used to create an ensemble of GCM networks. As a concrete example of this method we extract the joint degree sequences using the SED, EECC and the MPCC of the GCC of the network science authorship network [50] in Fig 6.6 and Fig 6.7. Plotted are the experimental results from the original network (red

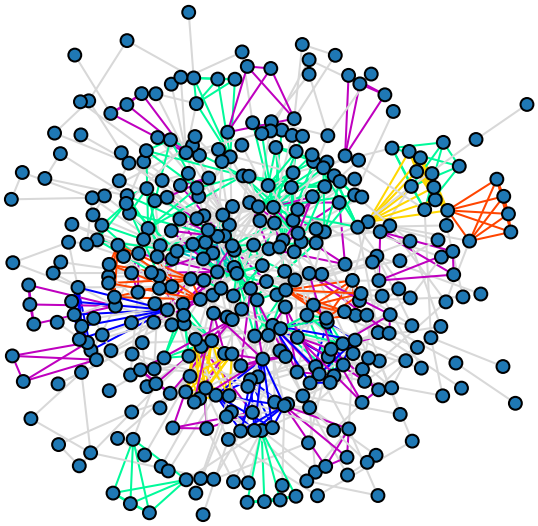


Figure 6.6: A member of the MPCC random graph ensemble of the GCC of the network science authorship network with higher-order cliques (larger than 3-cliques) coloured for clarity. Unlike random graphs constructed using the EECC method, larger cliques are preferentially retained in the ensemble.

crosses), the SED (green squares), the EECC (pink triangles) and the results from the MPCC algorithm (light blue) as well as their average (dark blue). The average neighbour degree, k_1 obtained from the SED shows poor accuracy when compared to the experimental results. Instead of the detailed NNDC structure over the range of focal vertex degrees, the neighbour degrees tend to fluctuate around $k_1 = 8$. In contrast, the MPCC exhibits a rich correlation structure whose average follows the trends of the experimental data. Additionally, the average neighbour degree for the high-degree vertices is well represented; however, this is at the expense of the lower degree information, where the representation is less accurate. The EECC shows fair agreement across the range of focal vertex degrees, outperforming the MPCC at low degrees. We notice from the variance of the MPCC that the NNDC of the empirical network is dense within the set of ensemble representations.

6.1.6 Anticorrelated modular networks

As an example of the generating function method, consider a multiplex network with anticorrelated clustering. Multiplex networks are a special class of multilayered networks [30] in which a set of vertices is connected by M different sets of coloured edges. Each layer contains a replicated set of vertices and connects them together with edges of a given colour. In anticorrelated networks, if a vertex has an edge of a given colour, then it has a vanishingly small probability of having edges of other colours. In this model, we extend that property beyond ordinary edges to the anticorrelation of the subgraphs that each vertex can belong to. The joint distribution for a maximally anticorrelated degree sequence, where the set of motifs in the model is $\boldsymbol{\tau} = \{1, \ldots, n\}$, is given by

$$p(k_1, \ldots, k_n) = \frac{1}{N} \sum_{v \in \tau} N_v p(k_v) \prod_{\omega \in \boldsymbol{\tau} \setminus \{v\}} \delta_{k_\omega, 0} \tag{6.41}$$

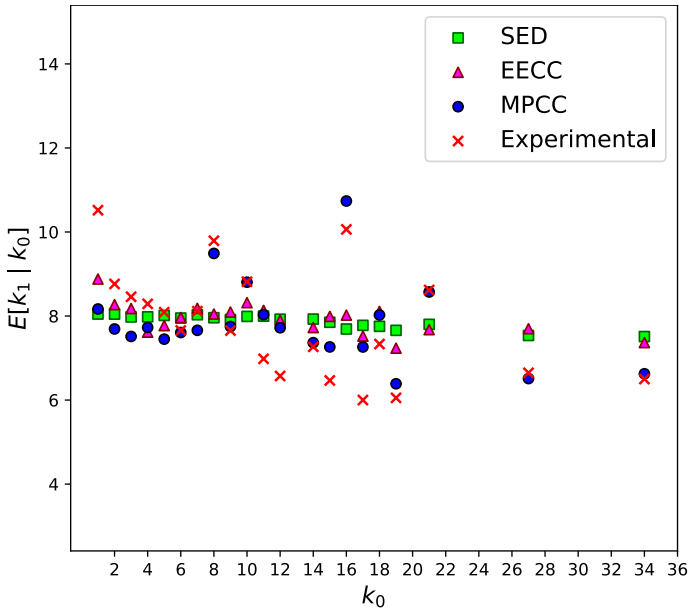


Figure 6.7: The ensemble expectation value of the overall degree of a neighbour as a function of focal vertex degree for clique covers of the network science authorship network. Plotted are the experimental results (red crosses), the average EECC (pink triangles), the average MPCC (dark blue circles) and its variance (light blue circles) for each realisation. Each simulation was performed 1000 times. The SED (green squares) doesn't capture the correlation structure for this network. The MPCC accurately captures the correlation structure of the high-degree vertices due to retaining the larger motifs that a vertex belongs to; however, the low (mid) degree sites are generally under (over) predicted. Conversely, the EECC performs well for the low and mid-degree vertices, but tends to the SED for the high-degree sites.

where $\delta_{i,j}$ is the Kronecker delta. For each $\nu \in \boldsymbol{\tau}$, k_ν is only non-zero when each k_ω for $\omega \in \boldsymbol{\tau}\backslash\{\nu\}$ is zero. Note N_ν is the number vertices in the network that are involved in topology ν.

When the marginal distribution in each ν is Poisson distributed with mean degree λ_ν, then we have

$$p(k_1,\ldots,k_n) = \frac{1}{N}\sum_{\nu=1}^{n} N_\nu \frac{\lambda_\nu^{k_\nu} e^{-\lambda_\nu}}{k_\nu!} \prod_{\omega\in\boldsymbol{\tau}\backslash\{\nu\}} \delta_{k_\omega,0} \tag{6.42}$$

Thus, each module is a component comprised of a given subgraph topology where motif membership is Poisson distributed about an average. The generating function for the probability of choosing a vertex at random from the network with a given degree sequence

is

$$\begin{aligned} G_0(z_1,\ldots,z_n) &= \frac{1}{N}\sum_{k_1=0}^{\infty}\cdots\sum_{k_n=0}^{\infty}\left(\sum_{\nu=1}^{n} N_\nu \frac{\lambda_\nu^{k_\nu}e^{-\lambda_\nu}}{k_\nu!}\prod_{\omega\in\boldsymbol{\tau}\setminus\{\nu\}}\delta_{k_\omega,0}\right) z_1^{k_1}\cdots z_n^{k_n} \\ &= \frac{1}{N}\sum_{k_1,\ldots,k_n=0}^{\infty}\left(\sum_{\nu=1}^{n} N_\nu \frac{\lambda_\nu^{k_\nu}e^{-\lambda_\nu}}{k_\nu!}\prod_{\omega\in\boldsymbol{\tau}\setminus\{\nu\}}\delta_{k_\omega,0}\right)\prod_{\mu=1}^{n} z_\mu^{k_\mu} \\ &= \frac{1}{N}\sum_{\nu=1}^{n} N_\nu \sum_{k_\nu=0}^{\infty}\frac{\lambda_\nu^{k_\nu}e^{-\lambda_\nu}}{k_\nu!}z_\nu^{k_\nu} \end{aligned} \tag{6.43}$$

Due to the condition of maximal anticorrelation the distribution is separable and so this expression reduces to

$$G_0(z_1,\ldots,z_n) = \frac{1}{N}\sum_{\nu=1}^{n} N_\nu e^{\lambda_\nu(z_\nu-1)} \tag{6.44}$$

The model can be numerically solved by defining a variable, u_ν for each $\nu\in\boldsymbol{\tau}$, that describes the probability that site ν remains unattached to the GCC. Each of these variables satisfies a self consistent equation

$$u_\nu = \frac{N_\nu}{N}e^{\lambda_\nu(g_\nu-1)} \tag{6.45}$$

where g_ν is the probability that a vertex in site ν fails to become attached to the GCC. Once these variables are found, the percolation properties follow from $S = 1-G_0(u_1,\ldots,u_n)$, see Fig 6.8. Performing the linear stability analysis at $u_\eta = 1$ for each module, the percolation thresholds are given by

$$\left((\tau-1)\frac{\langle k_\tau^2\rangle}{\langle k_\tau\rangle}-\tau\right)\leq 0 \tag{6.46}$$

where τ is the length of the topological cycle and $\langle k_\tau\rangle$ is the average number of cycles a vertex connects to.

In Fig 6.8, in order to highlight the individual contribution each module makes to the overall GCC on the network, we have set the Poisson mean degree of each module to be an order of magnitude apart. This means that the orange triangles, with mean λ percolate first, while the orange tree-like module, with mean degree 0.1λ percolates next. The green tree-like edges follow with mean degree 0.01λ and finally, the fourth phase transition occurs when the green triangles, with mean degree given by 0.001λ, connect to the GCC. We observe a stepped phase transition for the network as each module connects together. Each colour splits into a double phase transition; with additional hyperfine splitting within each layer associated with the anticorrelation between subgraph topologies.

6.2 Chapter summary

This chapter has opened a discussion on the nearest neighbour correlations between vertices with a given joint subgraph degree. To investigate this, we derived a correlation function that evaluated the probability of a vertex with a given joint degree having a neighbour with another specified joint degree. We plotted this correlation function against Monte Carlo simulation and found excellent agreement. We then investigated the correlations in detail for tree-triangle networks as well as some limiting examples. In a second line of research

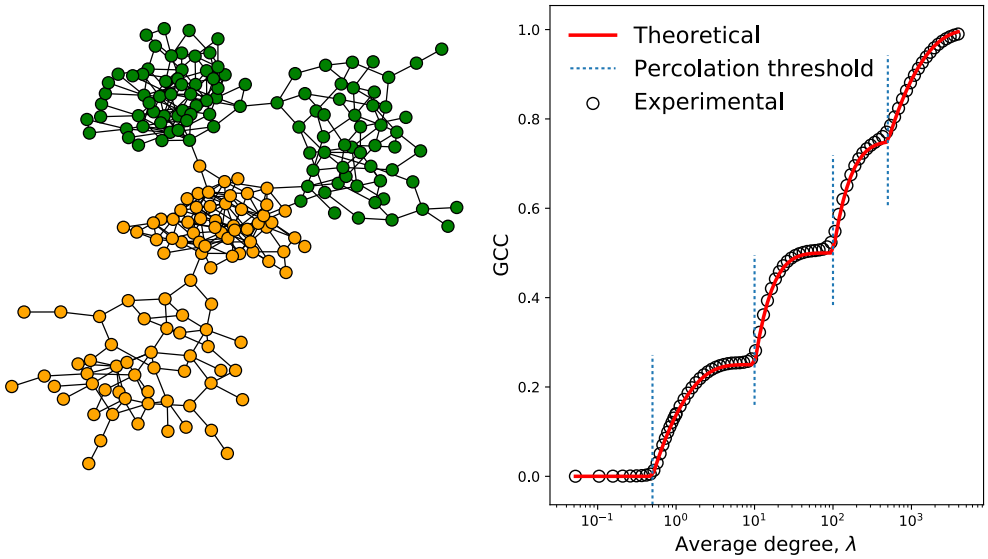


Figure 6.8: An example of an anticorrelated modular network with Poisson distributed subgraph membership in either 2- or 3-cliques coloured either orange or green and sparse inter-module connections (left). The percolation properties of the ensemble of such graphs (right). The average degree of orange triangles is given by λ; the average degrees of the remaining topologies have been set such that their percolation transitions occur orders of magnitude apart from one another. Specifically, the Poisson average of the number of orange tree degrees a vertex belongs to is 0.1λ, the green triangle average is 0.01λ whilst the green tree-like edges by 0.001λ. Vertical dashed lines indicate the predicted percolation threshold

we developed a novel stochastic clique covering algorithm which places edge-disjoint cliques over an empirical network. Our algorithm works by preserving the cliques with the largest size. If there are multiple cliques with a given size that overlap, then cliques are chosen at random to be part of the model. This enabled us to obtain the joint degree sequence for an empirical network and then, by constructing a joint degree distribution, enabled us to construct an ensemble of GCM networks. We then compared the properties of the ensemble, which we know that the theoretical correlation function supports, to the properties of the single-realisation empirical network. We also performed this for other clique covers in the literature. We displayed our results by compressing the joint degree information into an overall degree, finding fair agreement with the empirical network's overall degree correlations.

We also studied the percolation properties of multilayer anticorrelated networks with clustered topologies. We found that it was possible to isolate the phase behaviour of each module.

CHAPTER SEVEN

TWO-STAGE EPIDEMICS ON CLUSTERED NETWORKS

In this chapter we utilise the percolation-SIR equivalence to consider three 2-strain sequential epidemic processes on clustered networks using the generating function method. This defines a new field of investigation and is thematically distinct from the preceding chapters. In each of the models, the nature of the interaction between the diseases is changed; the first two are extensions of the framework Newman introduced [49, 53] for tree-like networks to the case of clustered networks. The third is a generalisation of these models also for clustered networks. In the first instance, we consider a perfect cross-immunity disease interaction; this assumes the second strain will only infect vertices that did not belong to the GCC of strain 1. In the second model, we examine perfect coinfection; which, constrains disease 2 to spread only on the GCC of the first disease (if one exists). The third model relaxes these constraints and allows the second disease to infect all vertices in the network regardless of their residence state within the GCC or RG. Such a model is known as a partial immunity interaction, although our model also encompasses partial coinfection. This generalisation adds more realism to the model and accounts for a far wider spectrum of disease interactions between the limiting-case logic of Newman's models.

7.1 Complete cross-immunity

In this section we extend Newman's model of perfect cross-immunity on tree-like networks [49, 26] to graphs with clustering. In the cross-immune model, a bond percolation process is run over a substrate network to create a GCC and a RG. The vertices in the RG are then percolated a second time by a second disease.

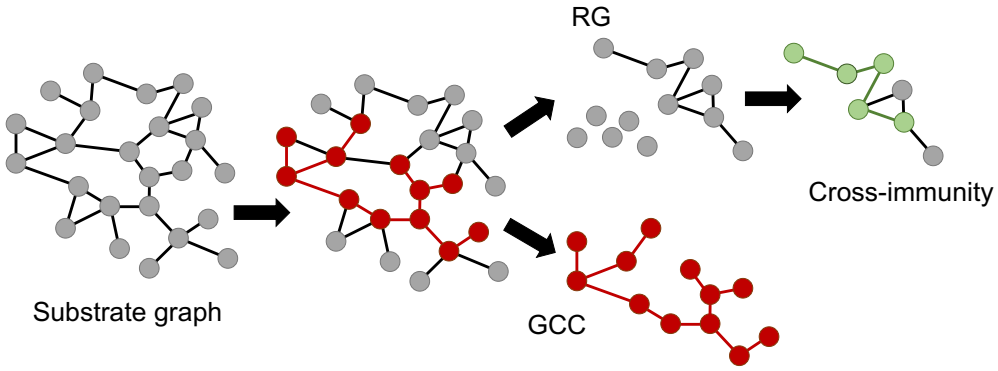


Figure 7.1: A conceptualisation of the cross-immunity model. A substrate network undergoes bond percolation to create a GCC and an RG. In the cross-immunity model, the RG is then percolated further by a second bond percolation process to create an embedded GCC (green). Vertices that were not in the RG of the first process cannot be included in the GCC of the second process.

By now, we are well aware of how to construct the probability that a randomly chosen vertex does not belong to the GCC in clustered networks. Restricting our attention to a mixed 2- and 3-clique GCM graph, the outbreak size of an epidemic is given by

$$S_1[u_2,u_3;T] = 1 - G_0(g_2, g_3^2) \tag{7.1}$$

where

$$u_\tau = G_{1,\tau}(g_2, g_3^2) \tag{7.2}$$

and with

$$g_2(u_2;T) = u_2 + (1-u_2)(1-T) \tag{7.3}$$

$$g_3^2(u_3;T) = \binom{n_3}{l}[u_3^2]^l \binom{\eta_3 - l}{m}[((1-u_3)(1-T))^2]^m \times [2u_3(1-u_3)(1-T)(1-T^2)]^{\eta_3 - l - m} \tag{7.4}$$

7.1.1 Strain-2

Once the first strain has passed through the network, a fraction, S_1, of the vertices will have contracted it and consequently a fraction, $1-S_1$, remained uninfected. In the case that vertices infected by strain 1 have perfect cross immunity against further strains, then only those vertices in the RG, $1-S_1$, can become infected by the second strain. The threshold criterion for the emergence of the second strain on unclustered random graphs has been

solved previously by Newman [49]. We now proceed to understand the role of clustering on the second strain.

Setting the transmissibility of the second strain to T_2, the probability that the second strain fails to infect a vertex chosen at random is comprised of the probabilities that both the tree-like edges and the triangle edges each fail to transmit the strain. In analogy to the first disease, we define the probability h_2 to be the probability that a tree-like edge remains unoccupied following both strains and introduce v_2 is the probability that a neighbouring vertex at the end of a tree-like contact does not have disease 2. The probability that a vertex with k tree-like contacts has precisely $l \leq k$ susceptible neighbours following disease 1 of which $m \leq l$ also failed to contract disease 2 is given by

$$h_2(u_2, v_2; T, T_2) = \binom{k}{l}\binom{l}{m}[u_2 v_2]^m [u_2(1-v_2)(1-T_2)]^{l-m}[(1-u_2)(1-T)]^{k-l} \quad (7.5)$$

Similarly, the probability, h_3^2, that a focal vertex involved in a triangle fails to become infected is given by the probability that each avenue of infection fails, as considered for the first disease in Eq 7.4. Defining v_3 to be the probability that a vertex involved in a triangle, that is also in the RG of the first strain, remains uninfected during the second epidemic, we now examine each bracket in Eq 7.4.

In the first case, both vertices are uninfected with strain-1 with probability u_3^2. To remain uninfected with strain-2, these vertices must fail to transmit to the focal vertex. This can occur in three distinct ways: either both neighbours fail to contract strain-2, v_3^2, or they both have disease-2 but fail to transmit, $((1-v_3)(1-T_2))^2$, or finally, one remains uninfected with strain-2 and the other fails directly to infect with probability $2v_3(1-v_3)(1-T_2)$.

Next, in the case when the RG contains both an infected and an uninfected vertex, there are only two ways that the focal vertex can remain uninfected by strain-2. These are the probability that the neighbour remains uninfected, v_3, or is infected but fails to transmit, $(1-v_3)(1-T_2)$. Together, these terms can be written as

$$\begin{aligned} h_3^2(u_3, v_3; T, T_2) = &\binom{\eta}{l}[u_3^2]^l \binom{l}{j}[v_3^2]^j \binom{l-j}{i}[2v_3(1-v_3)(1-T_2)(1-T_2^2)]^i \\ &\times [((1-v_3)(1-T_2))^2]^{l-j-i}\binom{\eta-l}{m}[2u_3(1-u_3)(1-T)(1-T^2)]^m \\ &\times \binom{m}{f}[v_3]^f[(1-v_3)(1-T_2)]^{m-f}[((1-u)(1-T))^2]^{\eta-l-m} \end{aligned} \quad (7.6)$$

Upon application of the binomial theorem this expression becomes

$$\begin{aligned} h_3^2(u_3, v_3; T, T_2) = &[u_3^2[v_3^2 + 2v_3(1-v_3)(1-T_2)(1-T_2^2) + [(1-v_3)(1-T_2)]^2] \\ &+ [2u_3(1-u_3)(1-T)(1-T^2)[v_3 + (1-v_3)(1-T_2)]] \\ &+ [((1-u_3)(1-T))^2] \end{aligned} \quad (7.7)$$

Despite the length of this equation, the interpretation is simple, we spread strain-2 according to the triangle formula of Eq 7.4 in the case that the residual motif is a complete triangle, we spread according to the tree-like expression when the residual triangle has only one neighbour in the RG; and finally, we do not spread strain-2 in the case that the motif is completely part of the GCC of strain-1. We can generate v_τ by writing self-consistent

expressions, this time however, dividing by the prior probability that the neighbour does indeed belong to the RG, which is simply u_τ.

$$v_\tau = G_{1,\tau}(h_2, h_3^2)/u_\tau \tag{7.8}$$

The expectation value for the probability that a randomly chosen vertex fails to be infected by either strain is

$$A = \frac{G_0(h_2, h_3^2)}{1 - S_1} \tag{7.9}$$

where we have divided by the prior probability of belonging to the RG of disease 1. The fraction of the RG that belongs to the outbreak of the second strain, the giant residual connected component (GRCC), is then given by

$$S_2[u_\tau, v_\tau; T, T_2] = (1 - A)(1 - S_1) \tag{7.10}$$

The complete prescription is as follows: we use Eq 7.2 to compute $u_\tau \; \forall \tau \in \boldsymbol{\tau}$, we can then use Eq 7.1 to compute the epidemic outbreak size of the first strain. With these ingredients we calculate $v_\tau \; \forall \tau \in \boldsymbol{\tau}$ using Eq 7.8 before finalising the calculation of the second outbreak fraction with Eq 7.10.

7.1.2 Numerical results

A numerical example of the both strains can be seen in plot (C) of Fig 7.2 for varying clustering coefficients. The networks for the model are created according to the configuration model [40, 52] where the stub-degrees of both tree-like (k_2) and triangle $(t = k_3/2)$ topologies of each vertex are Poisson distributed. The joint degree-distribution is given by [52]

$$p(s,t) = e^{-\mu}\frac{\mu^{k_2}}{k_2!}e^{-\nu}\frac{\nu^t}{t!} \tag{7.11}$$

where μ is the average tree-like degree and ν is the average number of triangles. The clustering of each network is varied such that the mean degree is fixed at 2. From this we find the means of each Poisson degree sequence as $\mu + 2\nu = 2$. As the clustering coefficient increases the epidemic threshold of the first strain decreases. Specifically, when $C = 0$ we have $\nu = 0$ indicating the threshold is $T_c = 1/2$, while at $C = 1/3$ we have $\mu = 0$ and hence find the critical threshold as the root of $T^2 + 2T - 1 = 0$ yielding $T_c \approx 0.41$.

The overall epidemic size at $T = 1$ is *reduced* as a function of increasing clustering coefficient. Therefore, in this experiment, clustering is seen to have a dual effect on the outbreak of strain-1 depending on T; clustered networks can expect an epidemic at lower T, but also expect fewer people to become infected. Setting $T_2 = 1$, the total outbreak size of the second strain decreases as a function of increased clustering.

In a second experiment we fix the degrees of each vertex according to the uniform-degree model, defined by Miller [40], enabling the effects of degree-assortativity to be understood. Bond percolation is run on three networks whose vertices have either degrees $2, 4$ and 6, but their clustering is distributed differently. The first has a joint degree distribution of $p(2,0) = 1/3$, $p(2,1) = 1/3$ and $p(0,3) = 1/3$, increasing the clustering of the high-degree sites. The second network has an even neighbour distribution with $p(2,0) = 1/6$, $p(0,1) = 1/6$, $p(2,1) = 1/3$, $p(4,1) = 1/6$ and $p(0,3) = 1/6$. Finally, the third network has clustering predominantly among the low-degree sites with $p(6,0) =$

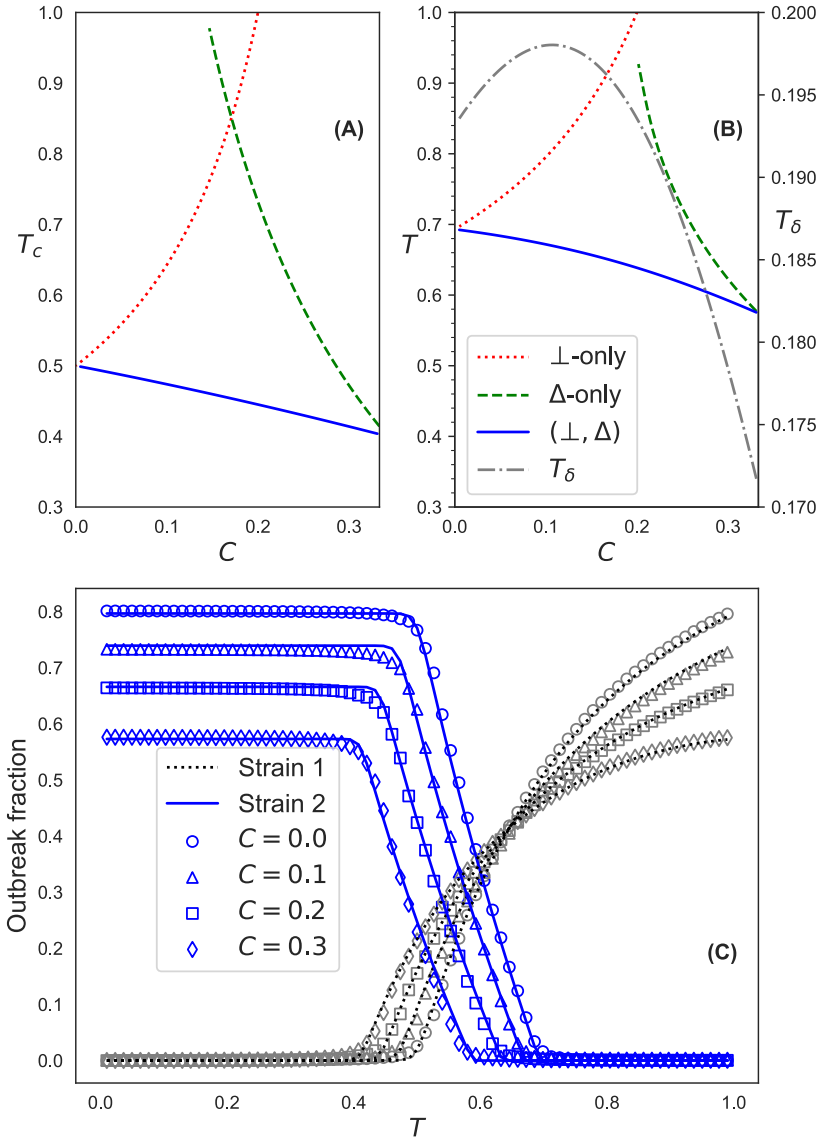


Figure 7.2: The percolation properties of the 2-strain model over clustered doubly-Poisson networks with clustering coefficient, C, and fixed average degree $\mu + 2\nu = 2$ of tree-like and triangles, respectively. (A) The epidemic threshold of strain-1 (solid) as a function of C. The critical thresholds for a GC to exist solely among tree-like edges (small dash) or triangle edges (long dash) from Eq 7.12 are plotted in (A). Similar analysis in plot (B) shows the coexistence threshold, T^*, as a function of increasing clustering coefficient. Also plotted in (B) is the difference $T_3 = T_c - T^*$ between the epidemic and coexistence thresholds. Plot (C) shows the expected epidemic size of each strain. Scatter points indicate experimental results of bond percolation on a network of size $N = 40000$ with 70 repetitions. Solid lines represent the theoretical predictions of Eqs 7.1 and 7.10 for each strain.

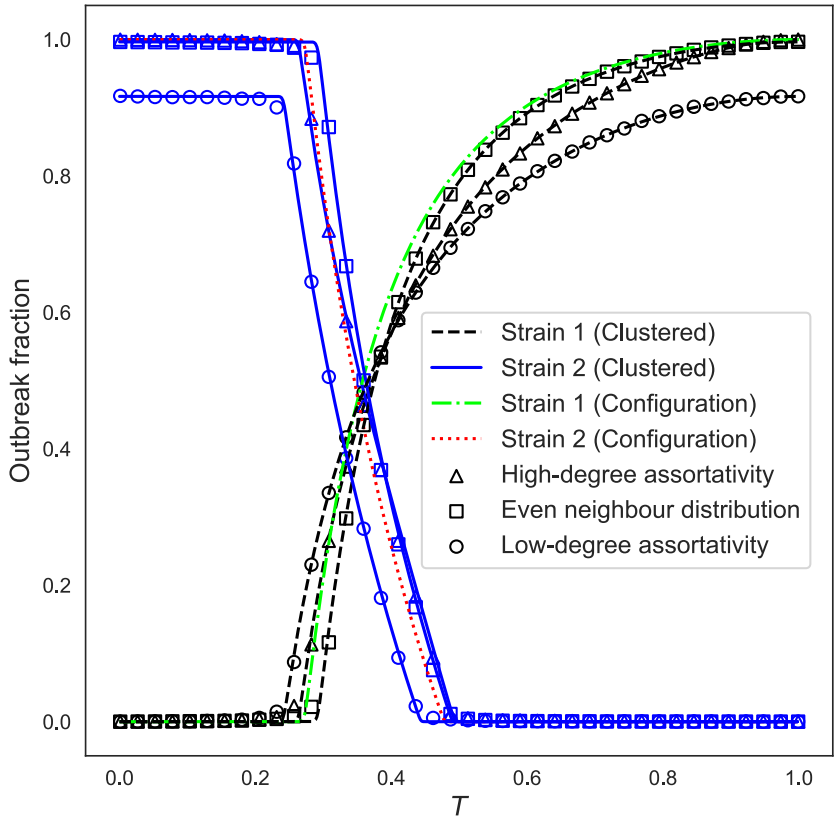


Figure 7.3: The residual network of the three networks described in section 7.1.1. In this model [40, 23], clustering can be shown to *increase* both the size of the GRCC and also the coexistence threshold relative to the configuration model. We also observe dichotomous results depending on the nature of the degree assortativity among the clustered edges. When clustering is assortatively confined to low-degree vertices, the results of the Poisson experiment are reproduced.

$1/3$, $p(2,1) = 1/3$ and $p(0,1) = 1/3$. The percolation properties of these networks are presented in Fig 7.3, along with the prediction from the configuration model. In contrast to the random Poisson networks, clustering is shown to increase both the GRCC and the coexistence threshold relative to the configuration model. Assortativity among low-degree clustered vertices leads to the emergent properties observed by the random Poisson networks.

7.1.3 R_0

The R_0 value, also known as the case reproduction number of a disease, is a quantity used in epidemiology to represent the number of infections that the average infected vertex in the network will cause. When the disease has a low transmissibility $T \leq T_c$, we do not expect that an epidemic will occur throughout the entire network, in other words, the

infections fizzle out over time. In these cases the R_0 value is less than unity. $R_0 = 1$ marks the threshold for which the epidemic infects a macroscopic fraction of the population and at this value the transmissibility experiences a critical point, $T = T_c$. Under the bond percolation isomorphism, a GC of occupied edges forms in the network at and after this bond occupancy probability. The critical transmissibility of the first strain can be found by applying the Molloy-Reed criterion to the configuration model [40]. Specifically, linearising $u_\tau = G_{1,\tau}(\boldsymbol{g}(\boldsymbol{u_\tau}))$ in ε around $u_\tau = 1 - \varepsilon_\tau$ [32]. To leading order in ε_τ we have $\boldsymbol{\varepsilon} = \boldsymbol{A\varepsilon}$ with $\boldsymbol{\varepsilon} = [\varepsilon_2, \varepsilon_3, \ldots]^{\mathrm{T}}$. The GC forms at the point when the determinant $\det|A - I|$ vanishes, where $A = [\partial \mathrm{G}/\partial u_\tau]$, $\mathrm{G} = [G_{1,2}, G_{1,3}]$ and I is the identity matrix. We thus obtain the following condition

$$\left(\frac{dg_2}{du_2}\frac{\langle k_2^2 - k_2\rangle}{\langle k_2\rangle} - R_0\right)\left(2\frac{dg_3}{du_3}\frac{\langle k_3^2 - k_3\rangle}{\langle k_3\rangle} - R_0\right) = 2\frac{dg_2}{du_2}\frac{dg_3}{du_3}\frac{\langle k_2 k_3\rangle^2}{\langle k_2\rangle\langle k_3\rangle} \tag{7.12}$$

where $\langle k_\tau\rangle$ is the first moment of the degree distribution (and similarly for other quantities) and each derivative is evaluated at the point $u_\tau = 1$. Each bracket on the left hand side can be used to investigate if a GC occurs among the edges of a given topology; or, the entire expression can be used to determine of the entire network is connected, irrespective of the edge-type, see plot (A) in Fig 7.2. It is clear from this plot that clustering increases the interval $T \in [T_c, 1]$ by the reduced epidemic threshold, allowing a finite-sized epidemic at lower transmissibilities.

Newman [49] found that the RG also experiences a phase transition due to the availability of vertices that are not within the GC as a function of T. In the case of clustered networks, we find the condition to be given by

$$\left(\frac{\partial h_2}{\partial v_2}\frac{\langle k_2^2 - k_2\rangle}{\langle k_2\rangle} - R_0\right)\left(2\frac{\partial h_3}{\partial v_3}\frac{\langle k_3^2 - k_3\rangle}{\langle k_3\rangle} - R_0\right) = 2\frac{\partial h_2}{\partial v_2}\frac{\partial h_3}{\partial v_3}\frac{\langle k_2 k_3\rangle^2}{\langle k_2\rangle\langle k_3\rangle} \tag{7.13}$$

The derivatives are evaluated at the point $v_\tau = 1$; however we must find the point (T^*, u_τ^*) that satisfies this where the *coexistence threshold*, T^*, signifies the emergence of a GC among the tree-like edges of the RG was derived previously by Newman [49].

As with the first strain, the presence of a GC of the second pathogen among only the tree-like or the triangle edges can be found by examining each bracket on the left hand side of Eq 7.13. The emergence of a GC among the entire RG is found using the entire expression, according to plot (B) in Fig 7.2. Setting $T_2 = 1$, we find

$$\partial_{v_2} h_2\big|_{v_2=1} = u_2^* \tag{7.14}$$

and hence the coexistence threshold among tree-like components is

$$T^* = \frac{u_2^* - 1}{G_{1,2}(u_2^*) - 1} \tag{7.15}$$

The coexistence threshold for the emergence of a GC among the triangles is slightly harder

to solve. Again, with $T_2 = 1$, we find

$$\partial_{v_3} h_3^2\big|_{v_3=1} = 2u_3^2 + 2u_3(1-u_3)(1-T)(1-T^2) \tag{7.16}$$

For brevity, we use the notation $\kappa = \langle k_3^2 - k_3\rangle / \langle k_3\rangle$ and hence we arrive at an equation just in T

$$(T^*)^3 - (T^*)^2 - T^* + \frac{2\kappa G_{1,3}(1,(g_3^*)^2)) - 1}{2\kappa G_{1,3}(1,(g_3^*)^2))(1 - G_{1,3}(1,(g_3^*)^2))} = 0 \tag{7.17}$$

where we have used Eq 7.2 to solve for u_3 given T in the absence of tree-like edges.

From plot (B) in Fig 7.2, it is clear that the interval $[0, T_1^*]$, which defines the transmissibility range within which strain-2 can exist on the network, is reduced as T^* decreases as a function of increasing C. Comparison of plots (A) and (B) indicates that while both T_c and T^* fall with C, the interval $[T_c, T^*]$, which defines the coexistence of each strain on the network, also is reduced, since, T^* falls faster than T_c. This indicates that clustering reduces the total fraction of the population affected at any given T; decreasing the range of values of T at which strain-2 can coexist with strain-1 present; and finally, decreasing the largest value of T at which strain-2 is found in the network, squeezing it to a smaller region of the model's phase space.

7.1.4 Cross-immune epidemics on multilayer networks

We will now apply the 2-strain model to clustered multilayer networks that exhibit modularity [56, 33]. For simplicity, we consider a 2-layer system comprised of tree-like edges in the first (orange) layer and triangle edges in the second (green) layer. In this example, the two layers are sparsely connected via interlayer tree-like edges; however, this is not a requirement, see Fig 7.4. Modular networks can be used to represent the different social contact structures that individuals might experience. For instance, a given family might have different contact topologies for schools, workplaces or social settings; each unique setting being represented by a distinct layer.

The multilayer model is an extension of the previous model; strain-2 spreading over the RG created by the GC of the bilayer networked system. Representing interlayer tree-like edges that an orange (green) vertex has as $\perp_{\rm og}$ ($\perp_{\rm go}$), the vector of permissible topologies is given by $\boldsymbol{\tau}_{\rm o} = \{\perp_{\rm o}, \perp_{\rm og}\}$ for the orange layer and $\boldsymbol{\tau}_{\rm g} = \{\Delta_{\rm g}, \perp_{\rm go}\}$ for the green layer, respectively, where $\Delta_{\rm g}$ represents a triangle in the green layer. Following [30, 33], each layer has its own $G_{0,\lambda}(\boldsymbol{z})$ equation, and each element of the topology vectors has its own $G_{1,\lambda,\tau}(\boldsymbol{z})$ equation also, where $\lambda \in \{{\rm o}, {\rm g}\}$ is a layer index.

As a numerical example consider the case where all edge topologies follow a Poisson distribution such that the number of τ edges is η_τ then

$$p_{\rm or}(\eta_\perp, \eta_{\perp,{\rm og}}) = \frac{\langle\eta_\perp\rangle^{\eta_\perp} e^{-\langle\eta_\perp\rangle}}{\eta_\perp!} \frac{\langle\eta_{\perp,{\rm og}}\rangle^{\eta_{\perp,{\rm og}}} e^{-\langle\eta_{\perp,{\rm og}}\rangle}}{\eta_{\perp,{\rm og}}!} \tag{7.18}$$

and

$$p_{\rm gr}(\eta_\Delta, \eta_{\perp,{\rm og}}) = \frac{\langle\eta_\Delta\rangle^{\eta_\Delta} e^{-\langle\eta_\Delta\rangle}}{\eta_\Delta!} \frac{\langle\eta_{\perp,{\rm go}}\rangle^{\eta_{\perp,{\rm go}}} e^{-\langle\eta_{\perp,{\rm go}}\rangle}}{\eta_{\perp,{\rm go}}!} \tag{7.19}$$

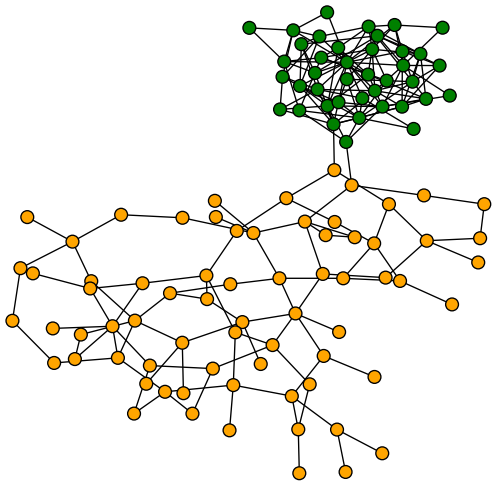


Figure 7.4: An example of the multilayer network used to in the numerical example. The green layer consists solely of triangles while the orange layer is tree-like. Each layer is connected via a few tree-like edges to allow the GC to span the network.

The expected outbreak size of the first epidemic on the orange layer is then

$$S_{\mathrm{o}} = 1 - e^{g_{\perp}(\langle \eta_{\perp} \rangle - 1)} e^{g_{\perp,\mathrm{og}}(\langle \eta_{\perp,\mathrm{og}} \rangle - 1)} \tag{7.20}$$

while the green layer has

$$S_{\mathrm{g}} = 1 - e^{g_{\Delta}(\langle \eta_{\Delta} \rangle - 1)} e^{g_{\perp,\mathrm{go}}(\langle \eta_{\perp,\mathrm{go}} \rangle - 1)} \tag{7.21}$$

The g_τ equations for each are given by Eqs 7.3 and 7.4 for the intralayer tree-like and triangle edges, respectively. The interlayer tree-like connections have a subtle symmetry breaking depending on which layer we consider the focal vertex to belong to. We define

$$g_{\perp,\mathrm{og}}(u_{\perp,\mathrm{go}}; T) = u_{\perp,\mathrm{go}} + (1 - u_{\perp,\mathrm{go}})(1 - T) \tag{7.22}$$

and

$$g_{\perp,\mathrm{g0}}(u_{\perp,\mathrm{og}}; T) = u_{\perp,\mathrm{og}} + (1 - u_{\perp,\mathrm{og}})(1 - T) \tag{7.23}$$

since, each focal vertex depends on the other end being uninfected. Each u_τ is then the solution to a self-consistent equation according to Eq 7.2.

The outbreak of the second epidemic follows from section 7.1.1 and in the Poisson case is

$$S_{2,\mathrm{o}} = 1 - e^{h_{\perp}(\langle \eta_{\perp} \rangle - 1)} e^{h_{\perp,\mathrm{og}}(\langle \eta_{\perp,\mathrm{og}} \rangle - 1)} \tag{7.24}$$

while the green layer has

$$S_{2,\mathrm{g}} = 1 - e^{h_{\Delta}(\langle \eta_{\Delta} \rangle - 1)} e^{h_{\perp,\mathrm{go}}(\langle \eta_{\perp,\mathrm{go}} \rangle - 1)} \tag{7.25}$$

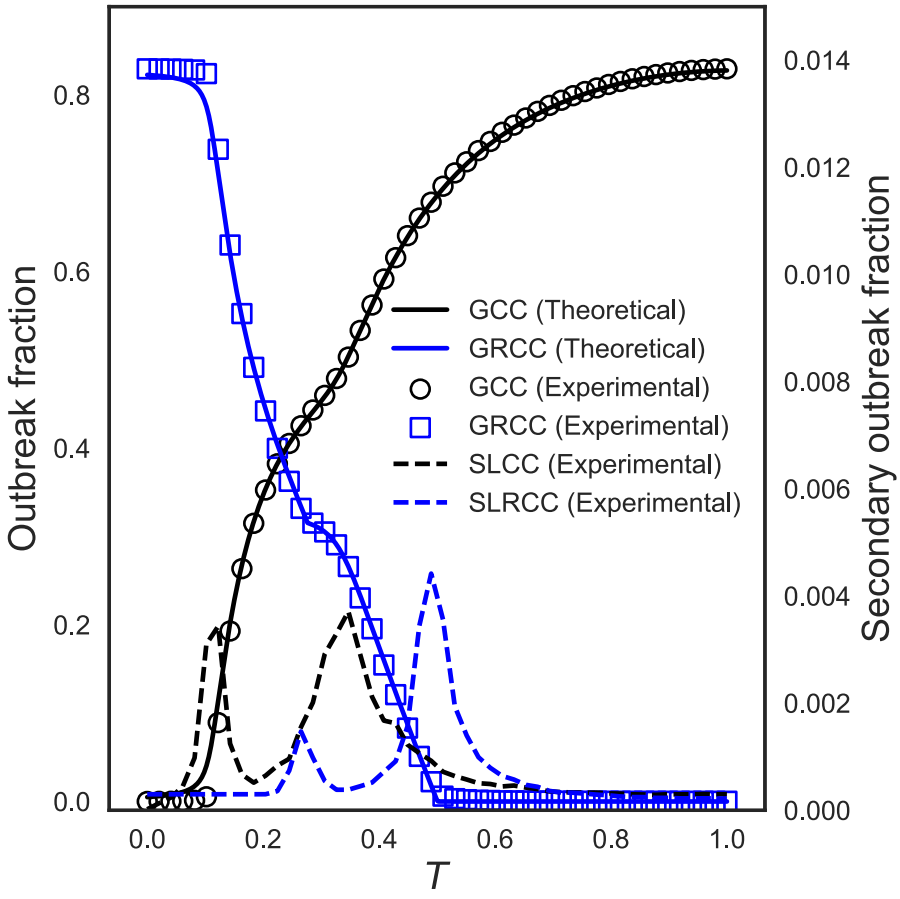


Figure 7.5: The expected epidemic size of each strain on a Poisson distributed clustered multilayer network with 2-layers (see Fig 7.4). In this experiment, the orange layer has a clustering coefficient of $C = 0$ while the green layer is set to $C = 1/3$. Interlayer tree-like edges have been added to allow the GC to span the entire network. Scatter points indicate experimental results of bond percolation on a network of size $N = 20000$ with 25 repeats. Solid lines represent the theoretical predictions of Eqs. Also plotted is the SLCC and the SLRCC, peaks in which indicate a phase transition. From this plot we can see that peaks in the SLCC and the SLRCC do not align with each other, their separation defines the region of coexistence of both strains.

We examine this system in Fig 7.5. The network is constructed such that the clustering coefficient of the green layer is $C = 1/3$ with mean degree $\langle k_\Delta \rangle = 6$ while the orange layer is $C = 0$ with mean tree-like degree $\langle k_\perp \rangle = 3.3$; a small number of interlayer edges were then added to connect the layers. In our experiment, the green-layer undergoes its phase transition at a lower T than the orange layer due to its clustering. This causes the outbreak fraction of the first strain to show a double 2nd-order transition [9, 33]. We confirm the presence of a phase transition by plotting the experimental second largest connected component (SLCC), peaks in which indicate a critical point.

Due to the different connectivity of each layer, the RG also experiences two critical points. We confirm this by plotting the second largest residual connected component (SLRCC), peaks in which indicate the presence of a phase transition in the residual network. The difference between the first peak in the SLCC and the last peak in the SLRCC defines the transmissibility range that allows coexistence of each strain in the network.

7.1.5 A model of the SLCC

In this section, we investigate how the two-pathogen model can be used to indicate the expected size of the second largest connected component (SLCC) for tree-like networks. We make the assumption that the giant connected component of the residual graph of a network that has undergone bond percolation (GRCC) is isomorphic to the SLCC. The model is based on [49] which studies the competition between two pathogens and can be adapted to the clustered model in this chapter by simply setting $T_2 = T_1$. In other words, the SLCC from the first percolation is assumed to have an equivalent structure to the GCC that can be formed from the RG (the GRCC), when the second strain percolates with occupancy equal to the first strain.

Let the probability that a neighbour is not connected to the GCC be u. Now, let v be the probability that the neighbour was not in the SLCC, given that it was not in the GCC.

We can generate a self-consistent expression for both u and v with the introduction of two generating functions

$$G_0(x) = \sum_k p(k)x^k, \qquad G_1(x) = \sum_k q(k)x^k \tag{7.26}$$

where $p(k)$ and $q(k)$ are the degree distribution and the excess degree distributions, respectively. It is well-known that we can write $u = G_1(1 - T + uT)$, finding a fixed point in $u \in [0, 1]$. The fraction of the network occupied by the GCC is then

$$\text{GCC} = 1 - G_0(1 - T + uT) \tag{7.27}$$

Let us consider a vertex of degree k. The probability that: a fraction m, of the neighbours are not in either the SLCC or the GCC is $[uv]^m$; that $l - m$ of the k vertices are not in the GCC, but *are* in the SLCC, but that they fail to attach the focal vertex to the SLCC is $[u(1-v)(1-T)]^{l-m}$, and that there are $k - l$ neighbours in the GCC is

$$P(\text{SLCC} \mid \text{not in GCC}) = \binom{k}{l}\binom{l}{m}[uv]^m[u(1-v)(1-T]^{l-m}[(1-u)(1-T)]^{k-l} \tag{7.28}$$

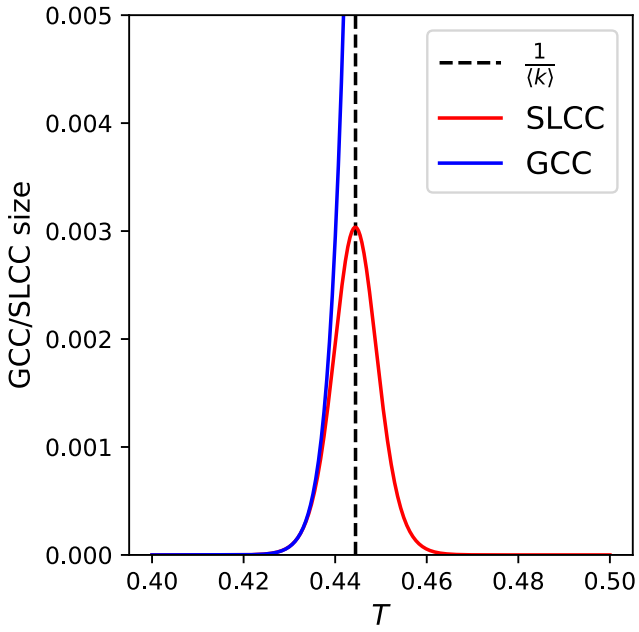


Figure 7.6: Theoretical curves of the GCC and the SLCC from Eqs 7.27 and 7.30, respectively, as well as the critical point (dashed) for Poisson networks with $\langle k \rangle = 2.25$.

Now, u is the solution to $u = G_1(1-T+uT)$, whilst v is the solution of

$$v = \frac{G_1(uv+u(1-v)(1-T)+(1-u)(1-T)}{G_1(1-T+uT)} \tag{7.29}$$

The size of the SLCC is then given by

$$\text{SLCC} = G_0(1-T+uT) - G_0(uv+u(1-v)(1-T)+(1-u)(1-T) \tag{7.30}$$

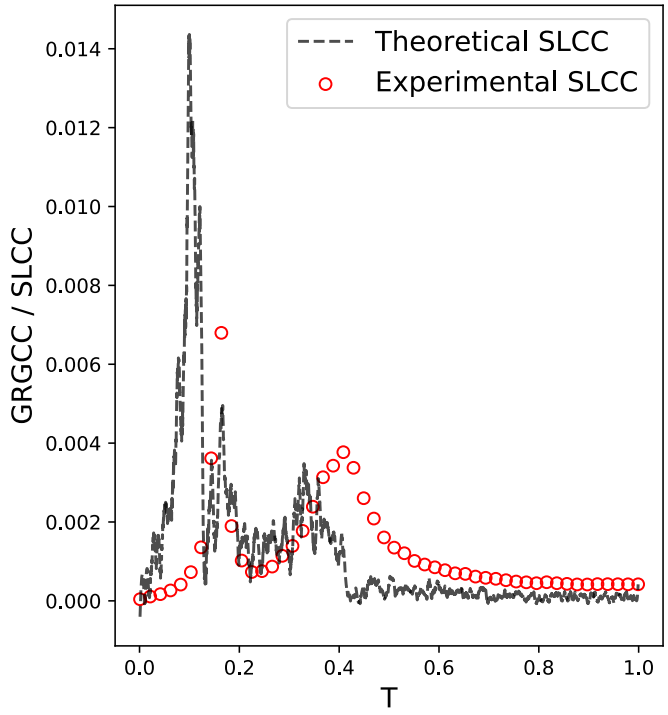


Figure 7.7: Plotted are the theoretical SLCC and the experimental SLCC fractions. The network is a a 2-layer network consisting of independent and triangular clustered edges according to the model in section 7.1.4. Percolation on this network was found to exhibit two critical points, corresponding to the emergence of a GCC on each layer. We see that the theoretical model correctly predicts two peaks in the SLCC. However, the theoretical line is noisy and not aligned with experiment.

7.2 Perfect coinfection

In this section we extend the model by Newman and Ferrario [53] to the realm of clustered networks. In a coinfection model, an epidemic occurs over a substrate graph. One equilibrium has been attained, the GCC created by the first process (if one is present) is then percolated a second time. In this paradigm, infection with the first strain is paramount for infection by the second. We examine the effects of clustering on the critical points as well as the outbreak sizes of coinfecting epidemics.

7.2.1 Strain-1

Once the dynamics of the first pathogen have run their course over the network, the vertices have either been infected or remain uninfected; a binary state equilibrium. To study an infected vertex in the GCC of strain-1, we must examine all the permissible final states that could surround the vertex through each edge type and assign a probability to each one. We can then sum the combinations of each state by creating a generating function that encapsulates the total probability of finding a particular infected vertex with a given final-state neighbour distribution. In this way, we use the local environment of the vertex to average over all possible neighbour states, weighted by the degree distribution, and then build a macroscopic description of the percolation properties of the entire network. To do

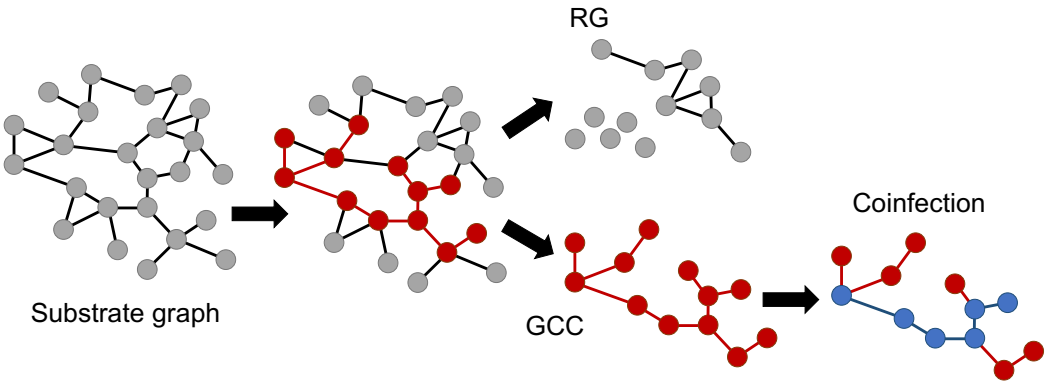


Figure 7.8: A conceptualisation of the coinfection model. A substrate network undergoes bond percolation to create a GCC and an RG. In the coinfection model, the GCC is then percolated further by a second bond percolation process to create an embedded GCC (blue). Vertices that were not in the GCC of the first process cannot be included in the GCC of the second process.

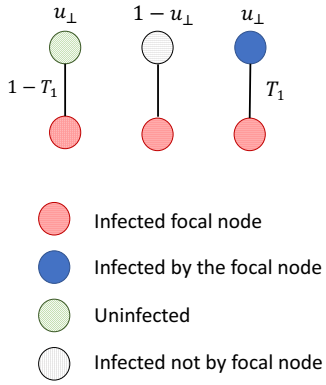


Figure 7.9: The tree-like edge topologies found in the GCC following the first strain and their probabilities with $u_\perp = u_2$.

this we must find the probability that the infected vertex transmitted or failed to transmit its infection to a neighbour during the dynamics of strain-1 through each topological edge-type. We do this first for tree-like edges as they are simpler than triangles and we note that a similar formula was found by Newman *et al* [53].

Assuming that the focal vertex is infected, there are three kinds of tree-like neighbours we can expect after strain-1: uninfected, infected (not by the focal vertex) and infected (by the focal vertex directly) according to Fig 7.9. Defining u_2 to be the probability that the neighbour found by following a tree-like edge was uninfected, the probability that it doesn't then become infected by the focal vertex is $1-T_1$. The probability that the neighbour was infected by vertices other than the focal vertex is simply $1-u_2$. Finally, vertices that were uninfected by their other neighbours can be infected directly by the focal vertex with probability u_2T_1.

Therefore, an infected vertex with s tree-like neighbours, of which l remain uninfected,

m are infected by their neighbours (other than the focal vertex) and $m' = s - l - m$ are infected directly by the focal vertex, occurs with the following probability

$$f_2(u_2;T_1) = \binom{s}{l}[u_2(1-T_1)]^l]\binom{s-l}{m} \times [1-u_2]^m[u_2T_1]^{s-l-m} \times [1-(1-T_1)^m] \tag{7.31}$$

The terminal bracket accounts for the probability that one of the m neighbours *must* have infected the focal vertex. This is expressed as one minus the probability that all m neighbours fail to infect it, each failure occurring with probability $1-T_1$.

The corresponding equation for triangles, $f_3^2(u_3;T_1)$ is a more involved calculation which we now examine. Defining u_3 to be the probability that a vertex involved in a triangle is uninfected, there are six basis triangles to consider following strain-1, see Fig 7.10.

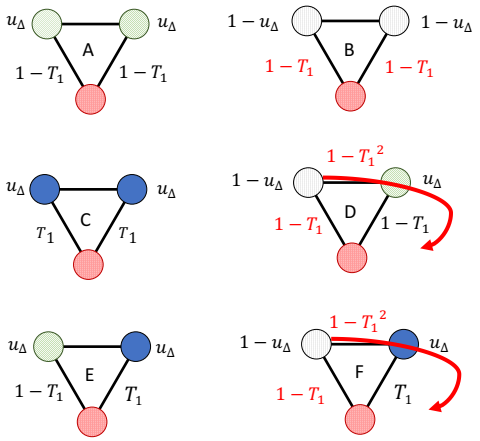


Figure 7.10: The 3-cycle edge-topologies found in the GCC following the first strain and their probabilities with $u_\perp = u_2$ and $u_\Delta = u_3$. vertex colours and patters are defined according to Fig 7.9. Triangles D, E and F consist of inhomogeneous neighbour-states and hence, due to the symmetry of the shape, their reflection about a vertical axis bisecting the focal vertex can also occur with equal probability. The curved arrows in triangles D and F indicate the additional pathway through the cycle that the infected neighbour could infect the focal vertex.

We will now discuss each triangle in turn from Fig 7.10. Triangle A assumes that both neighbours are uninfected, each with probability $u_3(1-T_1)$. Triangle B assumes that each neighbour is infected by means other than the focal vertex, each occur with probability $1-u_3$. Similarly to the tree-like edge-topology, the focal vertex infects a neighbour with probability u_3T_1, this occurs twice in triangle C. The remaining triangles D,E and F can also be formed by swapping each neighbour with equal probability of occurrence, hence, these configurations contribute twice to the neighbour-state distribution.

With these considerations in mind, the probability that a focal vertex involved in t triangles having precisely a of type A, b of type B (an so on) is

$$f_3^2(u_3;T_1) = \binom{t}{a}[u_3(1-T_1)]^{2a}\binom{t-a}{b}[1-u_3]^{2b}\binom{t-a-b}{c}[u_3T_1]^{2c}\binom{t-a-b-c}{d} \times [2u_3(1-u_3)(1-T_1)]^d\binom{t-a-b-c-d}{e}[2u_3^2T_1(1-T_1)]^e \times [2u_3(1-u_3)T_1]^{t-a-b-c-d-e}[1-(1-T_1)^{2b+d+e}(1-T_1^2)^{d+e}] \quad (7.32)$$

The terminal bracket accounts for the total probability that one of the infected neighbours (other than those the focal vertex infected) actually managed to transmit the infection to the focal vertex in the first place. This probability is constructed as one minus the probability that all the previously infected vertices failed to transmit their infection. Transmission can fail to occur in two ways in the 3-cycle: either directly with probability $1-T_1$, or around the cycle in the special case that the adjoining neighbour was initially uninfected, which occurs with probability $1-T_1^2$. Both methods are highlighted in red in Fig 7.10. Cycle B has two direct edges and cycles D and E are free to transmit around the outer skeleton of the triangle prior to the infection of the focal vertex.

Due to the terminal brackets in Eqs 7.31 and 7.32, the generating functions consist of two terms, the first considers all of the infections that all infected vertices create and amounts to unity, while the second subtracts those that were not part of the GCC or analogously the major outbreak of the strain due to the failure of the indirectly infected neighbours to infect the focal vertex.

To construct the generating functions we must insert both f_2 and f_3^2 into $G_0(x,y)$ and sum over each index

$$H_0(\vec{x}) = \sum_{s=0}^{\infty}\sum_{t=0}^{\infty} p(s,t) f_2 f_3^2 x_1^m x_2^{s-l-m} x_3^b x_4^c x_5^d x_6^e x_7^{\lambda-d-e} \quad (7.33)$$

with $\lambda = t-a-b-c$. Applying the binomial theorem we obtain

$$\begin{aligned} H_0(\vec{x}) = {} & G_0(u_2(1-T_1)+(1-u_2)x_1+u_2T_1x_2, u_3^2(1-T_1)^2+(1-u_3)^2x_3+(u_3T_1)^2x_4 \\ & +2u_3(1-T_1)(1-u_3)x_5+2u_3(1-T_1)u_3T_1x_6+2(1-u_3)u_3T_1x_7) \\ & -G_0(u_2(1-T_1)+(1-u_2)(1-T_1)x_1+u_2T_1x_2, u_3^2(1-T_1)^2 \\ & +((1-u_3)(1-T_1))^2x_3+(u_3T_1)^2x_4+2u_3(1-T_1)^2(1-u_3)(1-T_1^2)x_5 \\ & +2u_3(1-T_1)u_3T_1x_6+2(1-u_3)u_3T_1(1-T_1)(1-T_1^2)x_7) \end{aligned} \quad (7.34)$$

We will also define $H_{1,\tau}$ as the tautological analogue of Eq 7.34; however, in this case the G_0 generating function is replaced by $G_{1,\tau}$. Each u_τ value then satisfies a self-consistent equation given by

$$u_\tau = J_{1,\tau}(\vec{x}), \qquad \vec{x} = \{1,\ldots,1\} \quad (7.35)$$

where

$$\begin{aligned} J_{1,\tau}(\vec{x}) = {} & G_{1,\tau}(u_2(1-T_1)+(1-u_2)(1-T_1)x_1+u_2T_1x_2, u_3^2(1-T_1)^2 \\ & +((1-u_3)(1-T_1))^2x_3+(u_3T_1)^2x_4+2u_3(1-T_1)^2(1-u_3)(1-T_1^2)x_5 \\ & +2u_3(1-T_1)u_3T_1x_6+2(1-u_3)u_3T_1(1-T_1)(1-T_1^2)x_7) \end{aligned} \quad (7.36)$$

which is the argument of the second bracket of Eq 7.34. The outbreak size of the first

pathogen, S_1, can be generated by the following expression

$$S_1[u_2, u_3; T_1] = H_0(\vec{x}), \qquad \{x = 1,\ \forall x \in \vec{x}\} \tag{7.37}$$

7.2.2 Strain-2

When the second pathogen emerges on the network, the immunological landscape it experiences consists of vertices that became infected by the first disease and vertices that remained uninfected. An additional consideration is the infection history of each infected neighbour; they could have received the disease from the focal vertex itself or via other vertices they are connected to. To retain this important epidemiological information, it is necessary to define multiple probabilities, v_τ, of not becoming infected by the second disease for each scenario present in the GCC following the first pathogen. In other words, there is a v_τ value for each distinct vertex-site in Figs 7.9 and 7.10 that include disease-1 infected vertices. This is in analogy to the analysis in section 7.2.1 that defined a u_τ value for each subgraph that the first strain is incident upon in the contact network. In more detail, cycles B, C and F retain their triangle topology in the GCC of strain-1; cycles D and E have become fractured by strain-1 and hence spread strain-2 as if they were in fact trees; finally, the two neighbours in cycle A do not allow the proliferation of strain-2 under the limit of perfect coinfection because they both remain susceptible following the first epidemic. Given the topologies above, we determine that there are eight distinct vertex-sites and hence eight v_τ values required to define the second pathogen. These include: two tree-like values v_2^A and v_2^B for the externally- and directly-infected neighbours in Fig 7.9, respectively; along with v_3^B, v_3^C, v_3^D, v_3^E and v_3^{F1} and v_3^{F2} for each triangle in Fig 7.10 that has neighbours in the GCC of disease-1. Cycle F contains two infected vertices; however, their infection histories are non-equivalent, each requiring a unique description.

The probability, S_2, that the focal vertex does not contract disease-2, given that it had contracted disease-1 is the found to be

$$\begin{aligned} S_2 = H_0(&g(v_2^A), g(v_2^B), h^2(v_3^B), h^2(v_3^C), \\ &g(v_3^D), g(v_3^E), h^2(v_3^{F_1}, v_3^{F_2}))/S_1 \end{aligned} \tag{7.38}$$

where $g(v) = v + (1-v)(1-T_2)$ is the probability that a single edge fails to transmit disease-2; and

$$\begin{aligned} h^2(v_a, v_b) = &g(v_a)g(v_b) \\ &- (v_a + v_b - 2v_a v_b)T_2^2(1-T_2) \end{aligned} \tag{7.39}$$

with the notation convention that $h^2(v,v) := h^2(v)$, is the probability that infection fails when a vertex belongs to a triangle. The interpretation of Eq 7.38 is that the first bracket calculates the spreading of the second disease over the infected subgraph from which we then subtract those contributions that were not part of the GCC, or the major outbreak, of the network. It remains now to compute the 8 probabilities v_τ^α defined above. The recipe for these is quite straightforward: we compute the probability that each site fails to infect the focal vertex with disease-2, given that the focal vertex did indeed have disease-1. The $H_{1,\tau}(\vec{x})$ generating function can then be used to derive some of the probabilities that a neighbouring vertex belonging to a given site fails to infect the focal vertex. Within these 8 scenario probabilities, we must correctly normalise by the probability of obtaining a

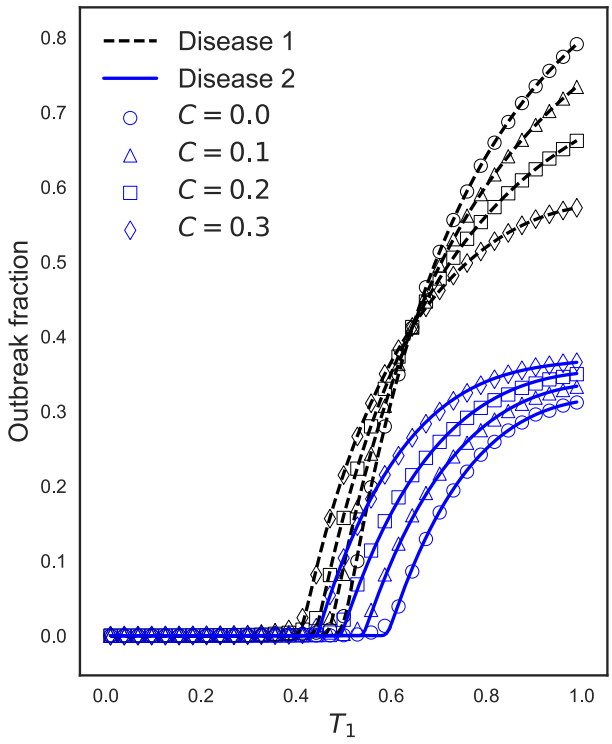


Figure 7.11: The outbreak fractions for the 2-strain coinfection model on doubly Poisson random graphs with $T_2 = 0.6$, $N = 25000$, for varying clustering coefficients and fixed overall mean degree $\langle s \rangle + 2\langle t \rangle = \langle k \rangle$ to be $\langle k \rangle = 2$. It is clear that clustering reduces the epidemic thresholds of both diseases and reduces the outbreak size of the first strain; however, it *increases* the coinfected fraction of the network. Scatter points are the average of 500 repeats of Monte Carlo simulation while solid lines are the theoretical predictions from Eqs 7.37 and 7.38.

particular site as well as multiplying by the probability that the focal vertex did or did not infect it. We find that

$$v_2^A = H_{1,2}(g(v_2^A), g(v_2^B), h^2(v_3^B), h^2(v_3^C), g(v_3^D), g(v_3^E), h^2(v_3^{F_1}, v_3^{F_2}))/(1-u_2) \tag{7.40a}$$

$$v_2^B = J_{1,2}(g(v_2^A), g(v_2^B), h^2(v_3^B), h^2(v_3^C), g(v_3^D), g(v_3^E), h^2(v_3^{F_1}, v_3^{F_2}))/u_2 \tag{7.40b}$$

$$v_3^B = H_{1,3}(g(v_2^A), g(v_2^B), h^2(v_3^B), h^2(v_3^C), g(v_3^D), g(v_3^E), h^2(v_3^{F_1}, v_3^{F_2}))/(1-u_3) \tag{7.40c}$$

$$v_3^C = J_{1,3}(g(v_2^A), g(v_2^B), h^2(v_3^B), h^2(v_3^C), g(v_3^D), g(v_3^E), h^2(v_3^{F_1}, v_3^{F_2}))/u_3 \tag{7.40d}$$

$$v_3^D = H_{1,3}(g(v_2^A), g(v_2^B), h^2(v_3^B), h^2(v_3^C), g(v_3^D), g(v_3^E), h^2(v_3^{F_1}, v_3^{F_2}))/(1-u_3) \tag{7.40e}$$

$$v_3^E = J_{1,3}(g(v_2^A), g(v_2^B), h^2(v_3^B), h^2(v_3^C), g(v_3^D), g(v_3^E), h^2(v_3^{F_1}, v_3^{F_2}))/u_3 \tag{7.40f}$$

$$v_3^{F_1} = H_{1,3}(g(v_2^A), g(v_2^B), h^2(v_3^B), h^2(v_3^C), g(v_3^D), g(v_3^E), h^2(v_3^{F_1}, v_3^{F_2}))/(1-u_3) \tag{7.40g}$$

$$v_3^{F_2} = J_{1,3}(g(v_2^A), g(v_2^B), h^2(v_3^B), h^2(v_3^C), g(v_3^D), g(v_3^E), h^2(v_3^{F_1}, v_3^{F_2}))/u_3 \tag{7.40h}$$

The complete prescription for solving the system of equations then involves: solving for the outbreak size of the disease-1 using Eqs 7.36 and 7.37 before solving the simultaneous system of Eqs 7.40a to 7.40h to find the v-values and finally using Eq 7.38 to find S_2. The fraction of the network that then contracts both diseases is then given by $S_1(1-S_2)$.

7.2.3 Numerical results

In this section we consider the coinfection model for two random graph ensembles. In each scenario, the model is compared with Monte Carlo simulation. The details of the simulation is as follows: a substrate network undergoes bond percolation at a given bond occupation probability T_1. The GCC is then percolated at a second occupation probability T_2 which represents the coinfected sub-population.

An example of the 2-strain model is shown in Fig 7.11 for networks with fixed mean degree $\langle k\rangle = \langle s\rangle + 2\langle t\rangle = 2$. In the limit of large network sizes, the joint degree distribution is given by the double Poisson distribution

$$p(s,t) = e^{-\langle s\rangle}\frac{\langle s\rangle^s}{s!}e^{-\langle t\rangle}\frac{\langle t\rangle^t}{t!} \tag{7.41}$$

In this instance, the generating functions $G_0(x,y) = G_{1,\perp}(x,y) = G_{1,\Delta}(x,y) := G(x,y)$ and we can write

$$G(x,y) = e^{\langle s\rangle(x-1)}e^{\langle t\rangle(y-1)} \tag{7.42}$$

The clustering coefficient for these graphs is then found to be

$$C = \frac{2\langle t\rangle}{2\langle t\rangle + \langle k\rangle^2} \tag{7.43}$$

We can use this expression to determine the tree degree and the average number of triangles per vertex for a fixed average degree, $\langle k\rangle$, along with $\langle k\rangle = \langle s\rangle + 2\langle t\rangle$. We use this

expression in Fig 7.12 to numerically investigate the epidemic thresholds of the two strains as a function of increasing clustering coefficient. An outbreak is considered to be an epidemic if the fraction of the network infected, $S(T_1)$, is larger than $\varepsilon = 1 \times 10^{-3}$, and hence, we can use Eqs 7.37 and 7.38 to approximate the epidemic threshold of each strain to sufficient accuracy. It is clear that increased clustering coefficients lead to a broadening of the region of the model's parameter space which can sustain coinfection at the increasing expense of the single strain equilibrium.

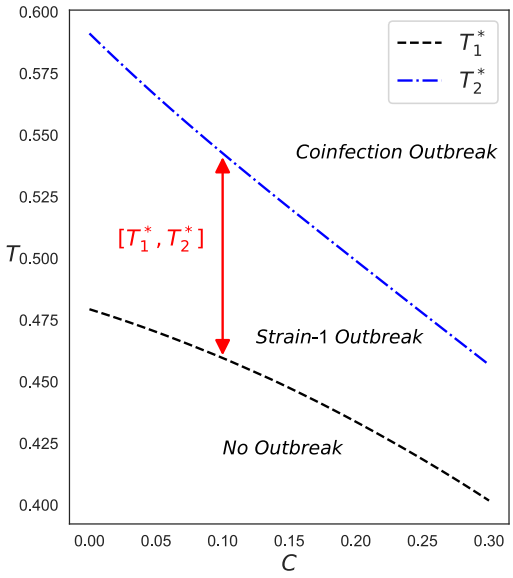


Figure 7.12: An analytical investigation of the critical transmissibilities of both strains for increasing clustering coefficients for the doubly Poisson networks defined in Fig 7.11. The curves are generated from Eqs 7.37 and 7.38 at the value of T in which the outbreak fraction becomes large than ε. The critical points reduce as C increases, indicating that contact clustering helps to spread an emergent epidemic. The interval $[T_1^*, T_2^*]$ is the transmissibility window in which strain-1 exists as an epidemic on the network without strain-2. We observe that increased clustering reduces the interval and thus increases the extent of coinfection in this model, at the expense of the mono-infection equilibrium. We note, however, that clustering can never reduce the coinfection critical point below the single strain threshold due to the strict conditions of perfect coinfection in the premise of the model.

The double Poisson model can display the properties of clustering as well as afford an exact solution. However, it is known that the distributions of contacts in many social networks are not well represented by the double Poisson distribution; instead, they often follow a power law. Further, whilst the average degree of Poisson networks can be fixed, the degree correlations are known to change with increasing C, which has been the subject of much research [40, 23, 20]. In this section, we investigate another clustering model that is more representative of real-world social networks as well as holding the degree correlations steady as we vary the clustering coefficient.

We propose a realistic model of human contact networks with tunable clustering based on Newman [46] and Hackett *et alia* [22, 20]. Contact networks often follow a scale-free (SF) distribution and of particular importance is the SF degree distribution, $p^{\mathrm{SFC}}(k)$, whose maximum degree is curtailed by an exponential degree cut-off (SFC) [10]. Such a distribution is given by

$$p^{\mathrm{SFC}}(k) = ck^{-\alpha}e^{-k/\kappa} \tag{7.44}$$

where $2 \leq \alpha \leq 3$ is the power law exponent, $\kappa \in \mathbb{Z}$ is the degree cut-off and c is a normalisation constant. In order to extend this degree distribution to the tree-triangle model, it is necessary to decompose the degree of a vertex, k, into tree-like and triangle contributions. We achieve this by introducing θ as the probability that a given vertex is a member of precisely t triangles. Thus, the clustered human contact network (CHCN) has tree-like and triangle degrees distributed according to

$$p^{\mathrm{CHCN}}(k) = p^{\mathrm{SFC}}(k) \sum_{t=0}^{\lfloor k/2 \rfloor} \binom{\lfloor k/2 \rfloor}{t} \theta^t (1-\theta)^{\lfloor k/2 \rfloor - t} \tag{7.45}$$

where $\lfloor \cdot \rfloor$ is the floor function. The normalisation constant can be found from the condition

$$\sum_{k=0}^{\infty} p^{\mathrm{CHCN}}(k) = 1 \tag{7.46}$$

Configuration model networks generated using this distribution have an identical distribution of overall degrees; however, their tree-like and triangle decompositions are modulated through θ. For high θ values, the heavy tail of the distribution is able to introduce a significant amount of clustering into the network. We observe the percolation properties of the model for increasing θ in the top plot of Fig 7.13. We find that clustering reduces the epidemic threshold of both strains and, in contrast to the double Poisson model, *reduces* the amount of coinfection in the network. Whilst CHCN networks have identical overall degrees, there is no control of the degree assortativity across the different experiments, however. Further, when the degree cut-off is large, we expect minimal assortativity, especially among the numerous low-degree sites. To examine the effect of assortativity on these results, we propose a use the degree-δ model [40] to the distribution to allow control of the degree correlations. The degree distribution is given by

$$p^{\mathrm{CHCN},\delta}(k) = \begin{cases} p^{\mathrm{CHCN}}(k)\delta_{k,s}\delta_{t,0} & k \neq 3 \\ p^{\mathrm{CHCN}}(k)\delta_{s,1}\delta_{t,1} & k = 3 \end{cases} \tag{7.47}$$

In other words, vertices in the network are not clustered unless their overall degree is $k = 3$, in which case, they are involved in exactly one triangle and one independent edge. This distribution forces the clustering to remain among the low-degree vertices towards the periphery of the network and thus, we expect clustering to be positively assorted. We examine this degree distribution in the bottom plot of Fig 7.13 for a lower degree cut-off and find, in agreement with [40, 20, 23], and in contrast to the results from Eq 7.45, that clustering *increases* the epidemic threshold of both strains.

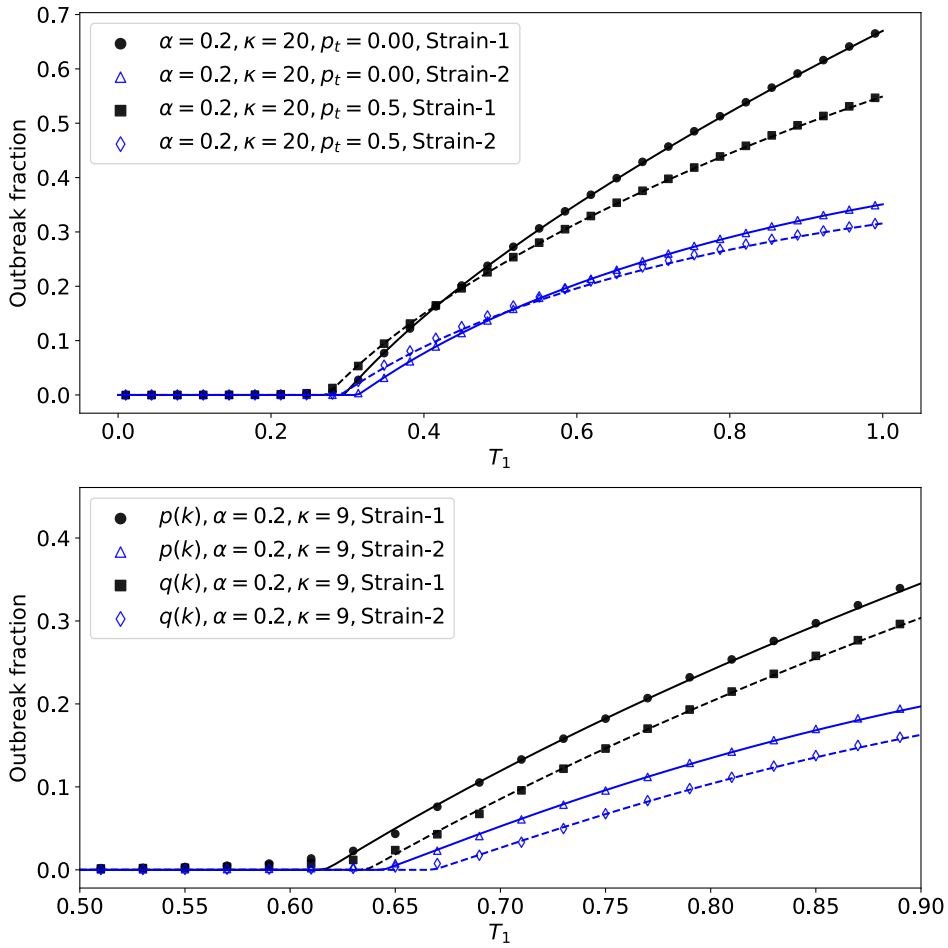


Figure 7.13: (Top) The outbreak fractions of both strains for increasing triangle probabilities within the CHCN degree distribution defined in Eq 7.45. Clustering reduces the epidemic threshold, in agreement with Poisson graph experiments; however, coinfection *decreases* with increasing θ. (Bottom) The expected epidemic sizes of both strains for a $p^{\mathrm{CHCN}}(k)$ network with $\theta = 0.0$ and a degree-δ CHCN distribution network defined by Eq 7.47. Clustering *increases* the epidemic threshold in this model. Markers are the average of 100 repetitions of bond percolation on CHCN networks with $N = 10,000$ vertices and $T_2 = 0.6$.

7.3 Partial immunity

In the previous section we extended Newman's work on temporally separated, two strain epidemics that are constrained to spread on either the RG [49] or the GCC [53] created by the preceding epidemic, on tree-like networks to the clustered network GCM paradigm [31, 34]. We then investigated the role that clustering plays on the epidemic properties, such as the critical points of the model and the size of the GCC. There are many avenues that we could explore from this starting point including: the effect of the presence of other subgraphs, such as cliques or chordless cycles; subsequent percolation events (which we do in chapter 8) or perhaps the most important generalisation, the relaxation of the strict conditions that the subsequent process is forced to spread under. In other words, a model in which the proliferation of the second disease is only *modulated* in some manner

(either competitively or cooperatively), but the infection state of a vertex is not a hard constraint on catching the second disease or not. Such a model is a *partial interaction model*; which we colloquially simplify to a *partial immunity* model despite this implying only competition [34]. The model is conceptualised in Fig 7.14 and is the subject of this chapter.

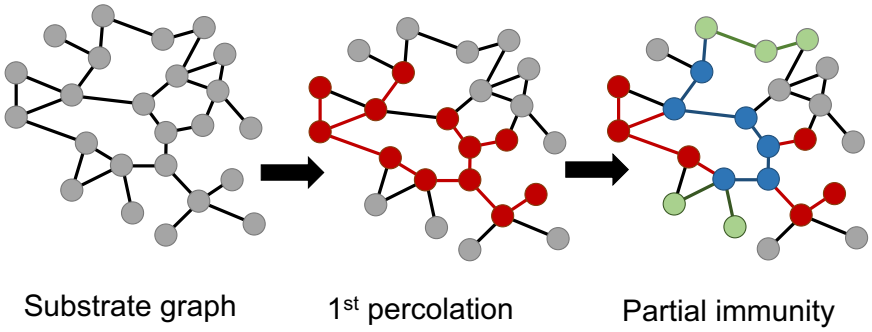


Figure 7.14: A conceptualisation of the partial immunity model. A substrate network undergoes bond percolation to create a GCC and an RG. In the partial immunity model, the entire network is then percolated further by a second bond percolation process to create its own GCC. All vertices in the network, despite their residence following the first percolation process, can be included in the GCC of the second process. Vertices that were in the GCC (RG) of the first process and are present in the GCC of the second process are coloured blue (green).

The defining feature of the partial immunity model lies in the discussion from section 2.6 regarding the different interpretations of g_2. Chapter 3.4, on the complement solution, also hints at this by writing the probability of belonging to the GCC in full rather than using the mutually exclusive "1 minus" approach.

Consider the neighbourhood of the two vertices that we might choose from the equilibrium of the first disease in a clustered network. Across both focal vertices, we observe 14 different motifs that could surround a pair of vertices chosen at random. Among these, there are 18 different neighbour states with unique infection histories comprised from three basis states of neighbour vertex: uninfected (green), infected externally (grey) and infected directly (blue). For example, consider the infected focal vertex (red) in Figure 7.15. There are 9 different motifs [F-N] that could potentially surround the focal vertex. Counting each tree-like neighbour and each vertex within a triangle (excluding the focal vertex itself) that is not related by symmetry to its neighbour, there are 12 different neighbouring sites; each site is occupied by one of three infection states: uninfected (green), externally infected (grey) and directly infected (blue).

7.3.1 Uninfected vertex description

The local environment of a vertex in the RG created by the first strain is considered here [35]. This accounts for all motifs that have the yellow focal vertex in Figure 7.15. The

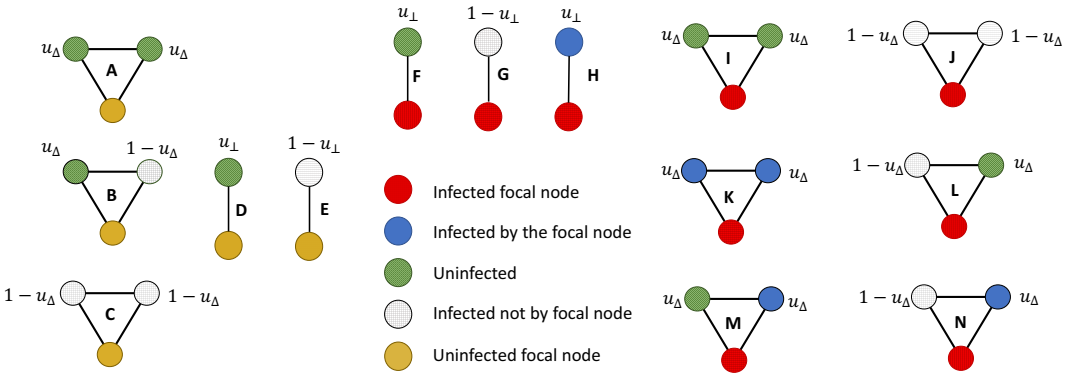


Figure 7.15: The 14 motifs that surround both focal vertices in the percolation model. In each case the lowest vertex in the motif is the focal vertex; of which there are two considerations, uninfected (yellow) and infected (red). There are three states a neighbouring vertex could be in: uninfected (green), infected externally (grey) and infected directly (blue), although the latter only occurs when the focal vertex is infected. Motifs [A-C] are the triangle motifs surrounding an uninfected focal vertex; D and E are the two types of tree-like edges. Motifs [F-H] are the tree-like edges that could surround an infected focal vertex and finally, motifs [I-N] are the triangles that an infected focal vertex can belong to. Among these motifs, there are 18 unique vertex-sites in total. Symmetric triangles (about a vertical axis bisecting the focal vertex) only contribute one site-type, whilst mixed triangles contribute two site-types and tree-like edges contribute one site-type each. The numbering convention for mixed triangles always proceed from left to right; for instance, in the mixed state triangle B the uninfected neighbour is B1 whilst the infected neighbour is B2. Also, we do not have to include the mirror image of mixed triangles, since, they occur with equal probability.

generating function for the probability of choosing an uninfected focal vertex (yellow) from the network, $p_{\text{uninfected}} = F_0(\vec{x})$, is

$$\begin{aligned} F_0(\vec{x}) =& G_0(u_2x_1 + (1-u_2)(1-T_1)x_2, (u_3x_3)^2 + ((1-u_3)(1-T_1)x_4)^2 \\ &+ 2u_3(1-u_3)(1-T_1)(1-T_1^2)x_5) \end{aligned} \tag{7.48}$$

where u_3 is the probability that a neighbour vertex in a triangle is uninfected. The vector $\vec{x} = \{x_1, \ldots, x_5\}$ has 5 dimensions, one for each of the 5 motifs that could surround the uninfected focal vertex. We define two additional generating functions $F_{1,2}(\vec{x})$ and $F_{1,3}(\vec{x})$ by replacing $G_0(\vec{x})$ in Eq (7.48) with $G_{1,2}(\vec{x})$ and $G_{1,3}(\vec{x})$, respectively. Eq (7.48) can be used to generate the size of the GCC of strain 1 according to the following self-consistent set of equations

$$u_2 = F_{1,2}(\vec{1}) \tag{7.49}$$

$$u_3 = F_{1,3}(\vec{1}) \tag{7.50}$$

followed by $S_1 = 1 - F_0(\vec{1})$.

7.3.2 Infected vertex description

The local environment of a vertex in the GCC created by the first strain is presented here [31]. The generating function, $p_{\text{infected}} = H_0(\vec{y})$, for picking an (externally) infected focal vertex (red) from the network that is part of the GCC is given by

$$\begin{aligned} H_0(\vec{y}) = G_0(&u_2(1-T_1)y_1 + (1-u_2)y_2 + u_2T_1y_3, (u_3(1-T_1))^2y_4 + (1-u_3)^2y_5 + (u_3T_1)^2y_6 \\ &+ 2u_3(1-T_1)(1-u_3)y_7 + 2u_3(1-T_1)u_3T_1y_8 + 2(1-u_3)u_3T_1y_9) \\ &- G_0(u_2(1-T_1)y_1 + (1-u_2)(1-T_1)y_2 + u_2T_1y_3, (u_3(1-T_1))^2y_4 \\ &+ ((1-u_3)(1-T_1))^2y_5 + (u_3T_1)^2y_6 \\ &+ 2u_3(1-T_1)^2(1-u_3)(1-T_1^2)y_7 + 2u_3(1-T_1)u_3T_1y_8 \\ &+ 2(1-u_3)u_3T_1(1-T_1)(1-T_1^2)y_9) \end{aligned} \tag{7.51}$$

We will also define $H_{1,2}(\vec{y})$ and $H_{1,3}(\vec{y})$ by replacing $G_0(\vec{y})$ by $G_{1,2}(\vec{y})$ and $G_{1,3}(\vec{y})$, respectively. Additionally, we generate a description of the directly infected neighbour state (blue) as

$$\begin{aligned} J_{1,\tau}(\vec{y}) = G_{1,\tau}(&u_2(1-T_1)y_1 + (1-u_2)(1-T_1)y_2 + u_2T_1y_3, (u_3(1-T_1))^2y_4 \\ &+ ((1-u_3)(1-T_1))^2y_5 + (u_3T_1)^2y_6 + 2u_3(1-T_1)^2(1-u_3)(1-T_1^2)y_7 \\ &+ 2u_3(1-T_1)u_3T_1y_8 + 2(1-u_3)u_3T_1(1-T_1)(1-T_1^2)y_9) \end{aligned} \tag{7.52}$$

The size of the GCC of strain 1 can be found by solving

$$u_2 = J_{1,2}(\vec{1}) \tag{7.53a}$$

$$u_3 = J_{1,3}(\vec{1}) \tag{7.53b}$$

and then $S_1 = H_0(\vec{1})$. In relation to the uninfected vertex description we have that $F_{1,\tau}(\vec{1}) = J_{1,\tau}(\vec{1})$ and that $H_0(\vec{1}) = 1 - F_0(\vec{1})$. Thus, the full description of the binary state equilibrium following bond percolation is given by the relation

$$1 = F_0(\vec{1}) + G_0(\vec{1}) \tag{7.54}$$

This expression constitutes a novel way, to our knowledge, of using the generating function formulation and it is this key equation that allows us to create the partial immunity model. This concept has been developed earlier in chapter 3.4, where the connections to the GCC were written explicitly in that case.

7.3.3 Strain 2

We have seen above how the GCC of the first strain can be obtained from either description of members of the percolation equilibrium: an uninfected vertex in the RG or an infected vertex in the GCC. To calculate the outbreak size of strain 2, we proceed as follows. For each of the 18 possible neighbouring vertex states, we must introduce a probability that infection with strain 2 does not occur through this channel by some means. Therefore, we introduce 18 distinct probabilities that a neighbour of a given state fails to infect a given focal vertex with strain 2. Although arbitrary, we choose different symbols for these probabilities depending on whether the neighbour state surrounds an uninfected vertex or

an infected vertex. We will see in a moment that subsets of the 18 sites are generated by the same expressions, and as such, the dimensionality of the model can be significantly reduced. However, we proceed in full for the moment.

There are 6 unique states that surround an uninfected focal vertex and thus, we define a set of 6 probabilities, $\{w\}$, that each hold the value of not becoming infected by strain 2 from one of these states. Specifically, there are 4 triangle neighbours and 2 tree-like neighbours so

$$\{w\} = \{w_3^A, w_3^{B1}, w_3^{B2}, w_3^C, w_2^D, w_2^E\} \tag{7.55}$$

Similarly, there are 12 states surrounding the vertex in the GCC and so we introduce a set, $\{v\}$, that holds the values of the probabilities of not becoming infected with strain 2 from these states. Specifically, there are three states reached by tree-like edges and 9 states within the triangle motifs. Hence,

$$\begin{aligned}\{v\} = \{&v_2^F, v_2^G, v_2^H, v_3^I, v_3^J, v_3^K, v_3^{L1}, v_3^{L2}, v_3^{M1},\\ &v_3^{M2}, v_3^{N1}, v_3^{N2}\}\end{aligned} \tag{7.56}$$

We next need to write self-consistent expressions for each of the values in $\{w\}$ and $\{v\}$. Before we do this, we define two functions that express the probability of transmission failing through a tree-like edge, $g(v,T) = v + (1-v)(1-T)$, and a triangle motif

$$\begin{aligned}h(v_\mu, v_\nu, T_\mu, T_\nu) = &\, g(v_\mu, T_\mu) g(v_\nu, T_\nu)\\ &- v_\mu (1-v_\nu)(1-T_\nu) T_\nu T_\mu\\ &- v_\nu (1-v_\mu)(1-T_\mu) T_\mu T_\nu\end{aligned} \tag{7.57}$$

with the convention that $h(v_\mu, v_\mu, T_\mu, T_\mu) = h(v_\mu, T_\mu)$. We will insert these functions into the $\vec{x}$ and $\vec{y}$ vectors in the arguments of the generating functions; each insertion describing the probability that strain 2 isn't contracted from a particular motif. The probability of not getting infected by strain 2 from the uninfected neighbour at the end of a tree-like edge is

$$w_2^D = F_{1,2}/u_2 \tag{7.58a}$$

The probability of not contracting strain 2 from the infected neighbour at the end of a tree-like edge is

$$w_2^E = H_{1,2}/(1-u_2) \tag{7.58b}$$

We now turn to the triangle probabilities $\{w_3^A, w_3^{B1}, w_3^{B2}, w_3^C\}$. The probability that the uninfected focal vertex doesn't get strain 2 from the symmetric susceptible site is

$$w_3^A = F_{1,3}/u_3 \tag{7.58c}$$

The probability that the symmetric infected site doesn't transmit to the uninfected focal vertex is

$$w_3^C = H_{1,3}/(1-u_3) \tag{7.58d}$$

The mixed triangle follows. For the uninfected focal vertex, we have the probability of not

becoming infected with strain 2 from an uninfected neighbour as

$$w_3^{B1} = F_{1,3}/u_3 \tag{7.58e}$$

Whilst for the infected site we have

$$w_3^{B2} = H_{1,3}/(1-u_3) \tag{7.58f}$$

We now have all of the probabilities that we require to describe the local environment of the uninfected vertex. We now turn to the description of the infected vertex in the GCC of strain 1. The three tree-like sites, $\{v_2^F, v_2^G, v_2^H\}$, are generated as follows: the uninfected neighbour

$$v_2^F = F_{1,2}/u_2 \tag{7.58g}$$

the externally infected neighbour

$$v_2^G = H_{1,2}/(1-u_2) \tag{7.58h}$$

and the directly infected neighbour

$$v_2^H = J_{1,2}/u_2 \tag{7.58i}$$

We now require the 9 triangle values $\{v_3^I, v_3^J, v_3^K, v_3^{L1}, v_3^{L2}, v_3^{M1}, v_3^{M2}, v_3^{N1}, v_3^{N2}\}$. The probability that an uninfected neighbour fails to transmit strain 2 through a symmetric uninfected triangle I is

$$v_3^I = F_{1,3}/u_3 \tag{7.58j}$$

The probability that the infected focal vertex in triangles J and K does not contract strain 2 is

$$v_3^J = H_{1,3}/(1-u_3) \tag{7.58k}$$

and

$$v_3^K = J_{1,3}/u_3 \tag{7.58l}$$

The mixed triangle L is given by

$$v_3^{L1} = H_{1,3}/(1-u_3) \tag{7.58m}$$

and

$$v_3^{L2} = F_{1,3}/u_3 \tag{7.58n}$$

Triangle M follows as

$$v_3^{M1} = F_{1,3}/u_3 \tag{7.58o}$$

and

$$v_3^{M_2} = J_{1,3}/u_3 \tag{7.58p}$$

Finally, triangle N is given by

$$v_3^{N_1} = H_{1,3}/(1-u_3) \tag{7.58q}$$

and

$$v_3^{N_2} = J_{1,3}/u_3 \tag{7.58r}$$

At this point, we have not yet written the arguments of each generating function, $\vec{x}$ and $\vec{y}$. It happens, that there are several equivalent expressions among the variables, allowing us to reduce the dimension of the problem considerably. Specifically, we notice the following redundancies among the relations: $v_3^{M1} = v_3^{L2} = v_3^I = w_3^{B1} = w_3^A$, $w_3^{B2} = w_3^C = v_3^{L1} = v_3^{N1} = v_3^J$, $w_2^E = v_2^G$, $w_2^D = v_2^F$, and $v_3^K = v_3^{N2} = v_3^{M2}$. This over prescription affords a reduction in the number of system variables to only 6 independent variables, one for each of the possible neighbour vertices: uninfected, externally infected and directly infected for tree-like edges and triangle motifs, respectively. Therefore, if we write the argument of each generating function $F_{1,2}, H_{1,2}, J_{1,2}, F_{1,3}, H_{1,3}$ and $J_{1,3}$ once, it is known for all occurrences of that function in the model. Further, we observe that the only difference between $F_{1,\tau}, H_{1,\tau}$ and $J_{1,\tau}$ for $\tau = 2,3$ is the underlying $G_{1,\tau}$ function, not the argument. In other words, the arguments of $F_{1,2}$ and $F_{1,3}$, for instance, are equivalent; we do not distinguish based on their topology. A final simplification can be achieved by noting that the arguments of $J_{1,\tau}$ and $H_{1,\tau}$ are also equivalent. Therefore, there are only two arguments to write: one for $F_{1,\tau}$ and another for $H_{1,\tau}$. These are given by $\vec{x} = \vec{\zeta}$ and $\vec{y} = \vec{\xi}$ where

$$\vec{\zeta} = \{g(w_2^D, T_2), g(w_2^E, T_2'), h(w_3^A, T_2), h(w_3^C, T_2'), h(w_3^{B1}, w_3^{B2}, T_2, T_2')\} \tag{7.59}$$

and

$$\begin{aligned}\vec{\xi} = \{&g(v_2^F, T_2), g(v_2^G, T_2'), g(v_\perp^H, T_2'), h(v_3^I, T_2), h(v_3^J, T_2'), h(v_3^K, T_2'), h(v_3^{L1}, v_3^{L_2}, T_2', T_2),\\ &\times h(v_3^{M1}, v_3^{M2}, T_2, T_2'), h(v_3^{N_1}, v_3^{N_2}, T_2', T_2')\}\end{aligned} \tag{7.60}$$

which constitute vectors of probabilities that each neighbour site fails to transmit infection to the focal vertex (or connect it to the GCC). With this in place, we now have an expression for all of the required probabilities $\{w\}$ and $\{v\}$. The size of the second outbreak over the network is then found by solving

$$S_2 = [F_0(\vec{1}) + H_0(\vec{1})] - [F_0(\vec{\zeta}) + H_0(\vec{\xi})] \tag{7.61}$$

where $[F_0(\vec{1}) + H_0(\vec{1})] = 1$. Qualitatively, this expression is 1 minus the probability that a vertex obtains strain 2 from either uninfected or infected neighbours.

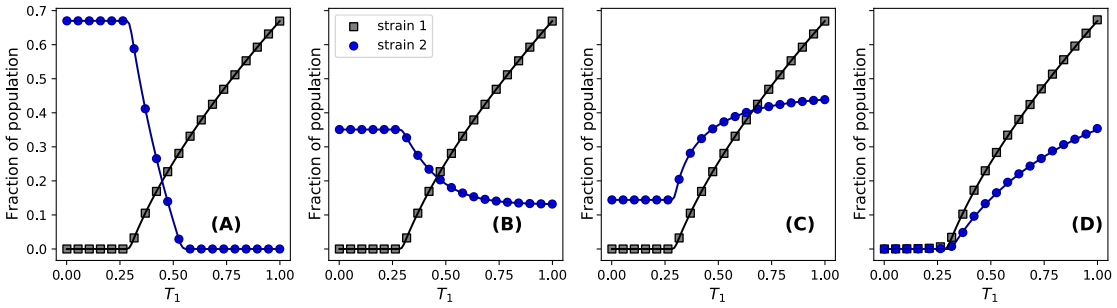


Figure 7.16: The outbreak fractions for several (T_2, T_2') combinations of the model all in the absence of clustering. From left to right: (A) complete cross-immunity $(T_2, T_2') = (1,0)$, (B) partial cross-immunity $(T_2, T_2') = (0.6, 0.39)$, (C) partial coinfection $(T_2, T_2') = (0.4, 0.7)$ and (D) perfect coinfection $(T_2, T_2') = (0, 0.6)$. Markers are the average of 50 repeats of bond percolation over CCM networks of size $N = 65000$, $\alpha = 2.0$ and $\theta = 0$; square markers are strain 1 whilst circles are strain 2. Solid lines are the theoretical results of Eq (7.61) for strain 2.

7.3.4 Numerical results

The results of the model under 4 different strain interactions for tree-like networks in the absence of clustering are shown in Figure 7.16 as T_1 is varied. Across the simulations, the networks are built according to the clustered contact model, (CCM) [31], which is an example of a configuration model degree distribution with clustering. Power law contact distributions are typical of those found in real-world social networks [48]. The underlying degree distribution is given by a power law model with exponential degree cut-off (PLC) defined as

$$p^{\mathrm{PLC}}(k) = \frac{k^{-\alpha} e^{-k/\kappa}}{\mathrm{Li}_{\alpha}(e^{-1/\kappa})} \tag{7.62}$$

where κ is the degree cut-off, $\alpha \in [2,3]$ is a power law exponent and $\mathrm{Li}_n(z)$ is the n-th polylogarithm of z [46]. Each k is then decomposed into tree-degrees, s and triangle degrees, t, according to Gleeson's method of edge partitioning [18]

$$p^{\mathrm{CCM}}(k) = p^{\mathrm{PLC}}(k) \sum_{t=0}^{\lfloor k/2 \rfloor} \binom{\lfloor k/2 \rfloor}{t} \theta^t (1-\theta)^{\lfloor k/2 \rfloor - t} \tag{7.63}$$

where $\lfloor \cdot \rfloor$ is the floor function and $\theta \in [0,1]$ is the probability of a pair of edges belonging to a triangle.

We simulate bond percolation for both strains numerically using Monte Carlo simulations. Following strain 1, infected vertices are labelled and subsequent infection with strain 2 occurs with probability T_2 for vertices in the RG or T_2' for vertices in the GCC.

In Figure 7.16a we have $T_2 = 1$ and $T_2' = 0$, a perfect cross-immunity model [49] in which infection with strain 1 prevents infection with strain 2. In Figure 7.16b we relax this hard limit, with $T_2 = 0.6$ and $T_2' = 0.39$, to obtain a partially cross-immune interaction whereby the transmission of strain 2 is reduced for strain 1 infected vertices. For $T_1 < T_{1,c}$ we observe the steady-state of strain 2 without competition from strain 1. At $T_1 = T_{1,c}$ a GCC in strain 1 emerges and the number of cases of strain 2 drops, but does not vanish;

in the limit $T_1 = 1$ strain 2 reaches its lowest incidence rate as competition is maximised. In Figure 7.16c we observe a partial coinfection model, with $T_2 = 0.4$ and $T_2' = 0.7$. In this case, strain 2 is facilitated by the presence of strain 1 in the network; the symbiotic interaction leading to an increase in the incidence of strain 2 infected vertices. Figure 7.16d shows the hard limit of a perfect coinfection model [53] with $T_2 = 0$ and $T_2' = 0.6$, strain 2 cannot survive without a GCC of strain 1 present in the network.

With an understanding of the model without clustering, we now examine the case where $\theta \neq 0$ for both partial interaction models with $\kappa = 20$, $\alpha = 2$ and $\theta = 0.5$, see Figure 7.17. The epidemic threshold of strain 1 is reduced with clustering, so too is the overall outbreak size of strain 1 at large T_1, compared to unclustered networks. The incidence of strain 2 exhibits dual behaviour over the range of T_1. For the partial immunity scenario (Figure 7.17a), with $T_2 = 0.6$ and $T_2' = 0.4$, clustering reduces the incidence of strain 2 at low T_1; however, it increases it as $T_1 \to 1$. Conversely, for partial coinfection (Figure 7.17b), with $T_2 = 0.4$ and $T_2' = 0.7$, having lowered the epidemic threshold of strain 1, clustering causes an increase in the incidence of strain 2 at lower T_1 values compared to the unclustered analogue.

In Figure 7.18 we perform a second experiment using the degree-δ model [40, 20, 31]. We define a distribution in which the degree of vertices involved in triangles is fixed to $k = 3$ and thus $(s,t) = (1,1)$, whilst all other degrees are given by Eq (7.63) for $(s,t) = (k,0)$. With the degree-correlations among triangles fixed, the epidemic threshold of the first strain increases with clustering. The partial cross-immune coupling (Figure 7.18a), with $T_2 = 0.8$ and $T_2' = 0.65$, no longer exhibits a cross-over in expected size of strain 1 and strain 2; clustering reduces the incidence of strain 2 for all values of T_1. Similarly, the partial coinfection model (Figure 7.18b), with $T_2 = 0.6$ and $T_2' = 0.75$, exhibits a reduced incidence of strain 2 compared to the unclustered analogue. As $T_1 \to 1$, however, the coinfection is reduced in the clustered graph compared to the unclustered.

7.4 Chapter summary

In this chapter we have introduced three models of 2-strain sequential epidemics on clustered networks. The first two are generalisations of Newman's prior work to the case of networks with finite measure clustering. These include the case of perfect cross-immunity and perfect coinfection, whereby, the second strain spreads solely on the RG or GCC, respectively. In the third model, we framed these two scenarios as limiting cases of a spectrum of infection criteria for a vertex to be infected with the second strain. Relaxing the infection prerequisites for the second epidemic allowed it to spread more readily over the network, with non-zero probabilities either side of the phase transition of the first strain.

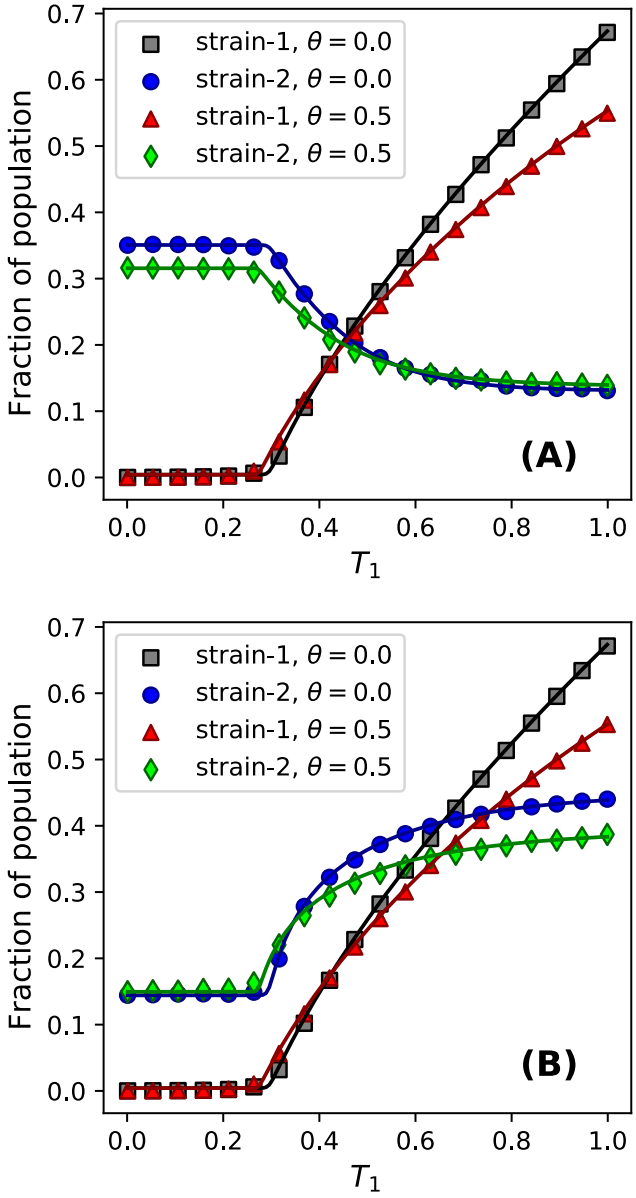


Figure 7.17: The outbreak fractions for both strains for clustered (Eq 7.63) and unclustered networks as T_1 is varied under two disease couplings: (A) partial cross-immunity $(T_2, T_2') = (0.6, 0.4)$ and (B) partial coinfection $(T_2, T_2') = (0.4, 0.7)$. Simulations are the average of 50 repeats of bond percolation on networks with $N = 35000$ and $\theta = 0.0$, $\alpha = 2.0$ and 0.5 for the unclustered and clustered networks, respectively. Solid lines are the theoretical results of Eqs 7.51 and 7.61. In general, clustering reduces the extent of plural infections in the network; however, degree assortativity within the contact topology causes a reversal of this at high (low) values of T_1 in A (B).

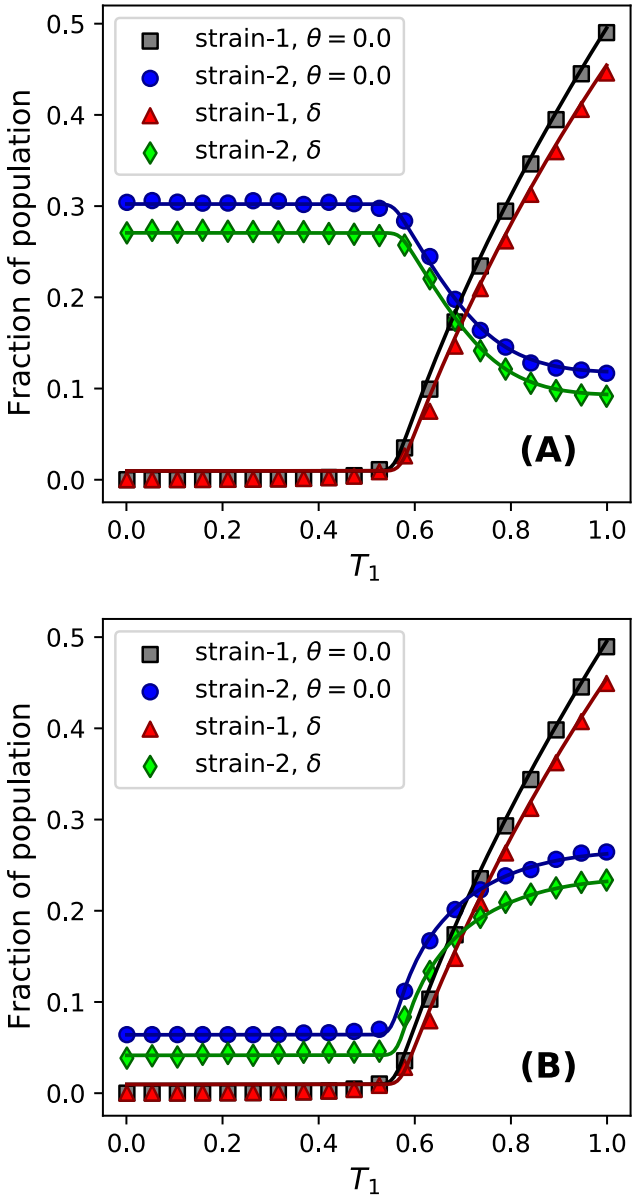


Figure 7.18: The outbreak fractions of the degree-δ model with clustering constrained to low degree assortativity for (A) partial immunity $(T_2, T_2') = (0.8, 0.65)$ and (B) partial coinfection $(T_2, T_2') = (0.6, 0.75)$ disease interactions. Simulations are the average of 50 repeats of bond percolation on networks with $N = 35000$, $\alpha = 2.0$; solid lines are the theoretical results of Eqs 7.51 and 7.61. Clustering reduces the fraction of the network that becomes infected by both strains for all values of T_1.

CHAPTER EIGHT

EPIDEMICS WITH *N*-STRAIN VARIANTS

In this chapter we formulate an extension of the epidemic models from the previous chapter by allowing additional generations of the percolation process to occur on the network. Specifically, we consider N generations of both the perfect cross-immunity model, which is manifestly a competitive process, and the perfect coinfection model, which conversely is a cooperative process. In the competitive process, subsequent percolation events occur on the RG created by the preceding generations, see fig 8.1 (top). In the cooperative process, future percolation events can only occur on the GCC created by the previous generations. We acknowledge that these models can be used to investigate the behaviour of repeated attacks on networks. For this purpose, we study the response of networks that have Poisson and scale-free degree distributions by examining the degree distribution and the cumulative degree distribution of the resulting graph structures. We find that scale-free networks are more robust than Poisson networks to repeated percolation.

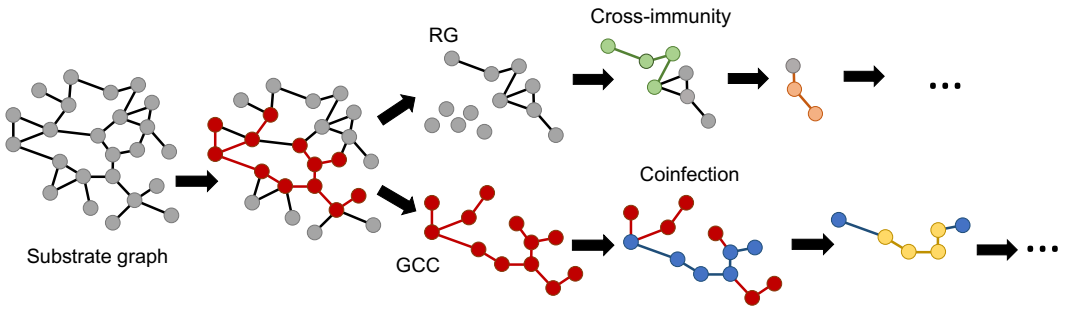


Figure 8.1: A conceptualisation of the both the cross-immunity (top) and coinfection (bottom) models with *N*-strain logic.

Whilst both of these branching processes are destined to burn out due to the exhaustion of available vertices, an N-strain partial interaction model would continue indefinitely as all vertices are potential substrates for future infection. We do not treat this case in the

thesis; however. We also restrict our attention to tree-like networks, rather than graphs with clustering.

8.1 N-strain cross-immunity

In this section, we define the i-th competitive branching process as successive bond percolations occurring on the RG created by the $i-1$ previous processes for $i = 1,\ldots,n$. This process has been studied previously using generating functions by Newman and Karrer when $n = 2$ [49, 26]. The structure of the RG has also been studied for clustered and modular networks [35]. From a network science perspective, this model allows us to study the structure of the RG of sequential bond percolation processes. In particular, we observe how those sequential processes fracture the RG into isolated components and study the phase behaviour associated with the sudden inability of the RG to support a GCC. Within the context of the SIR isomorphism, this model considers the behaviour of n seasonal strains of a disease (or separate diseases) that confer complete cross immunity to all subsequent pathogens. The model allows us to study the expected outbreak size of each generation and the point of natural burn-out due due to the shrinking of the susceptible sub-population.

8.1.1 Outbreak size

To the i-th process, $i = 1,\ldots,n$, we assign a bond occupation probability T_i and aim to calculate the probability that a randomly chosen vertex does not belong to a percolated component. From this, we can find the mutually exclusive probability that a vertex does belong to the i-th GCC, A_i. To do this, we define the probability that a neighbour of our randomly selected vertex is not part of the i-th GCC, u_i, given that it does not belong to any of the previous percolated components. Under the SIR isomorphism, T_i is the transmissibility of the i-th strain and u_i is the probability that a neighbour is not thus far infected.

There are two ways in which an edge emanating from the focal vertex can fail to connect it to the GCC: firstly, the neighbour could itself be unconnected, the probability of which by definition is u_i. Secondly, the neighbour could be connected, $(1-u_i)$, but the bond is unoccupied $(1-T_i)$. Therefore, the probability, $\bar{g}_i$, that an edge fails to connect the focal vertex to the i-th GCC given that the neighbour does not belong to any other GCC is

$$\bar{g}_i(T_i,u_i \mid \mathrm{RG}) = u_i + (1-u_i)(1-T_i) \tag{8.1}$$

The total probability that a neighbour belongs to the RG of the i-th percolation can then be found through a set of recursive functions, g_i that describe the probability that each iterative percolation failed to occupy this edge as

$$g_i(\boldsymbol{T},\boldsymbol{u}) = u_{i-1}\bar{g}_i + (1-u_{i-1})(1-T_{i-1}) \tag{8.2}$$

with $u_0 = 1$. A hierarchy of self-consistent equations [45] can be written to sequentially solve for each u_i value

$$u_i = \frac{G_1(g_i)}{\prod_j u_j}, \qquad j = 1,\ldots,i-1 \tag{8.3}$$

The size of the i-th GCC (epidemic) is then found by

$$A_i = \prod_{j=1}^{i-1} G_0(g_j) - G_0(g_i) \tag{8.4}$$

The total infected fraction of the network is given by $A = \sum_i A_i$.

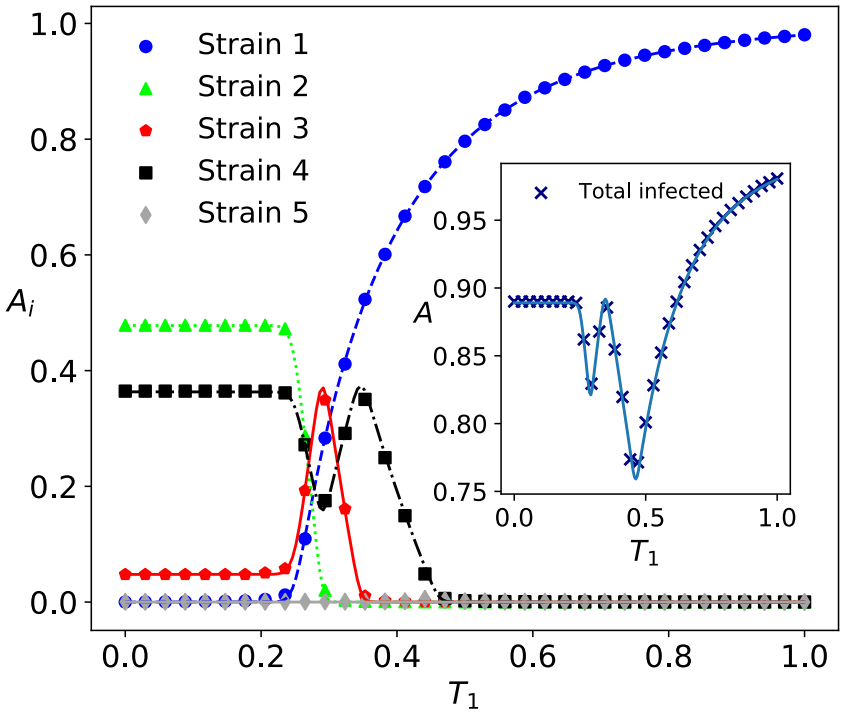


Figure 8.2: The outbreak fractions for five generations of the competitive branching process as a function of T_1. Solid lines are the theoretical results from Eq 8.4 whilst scatter points are the average of 50 repetitions of bond percolation over a network with $N = 35000$ vertices. The inset shows the total number of infected vertices.

As an example of Eq 8.4, we can obtain the expected outbreak size of the first epidemic, $n = 1$, from this system as $A_1 = 1 - G_0(g_1)$ where $u_1 = G_1(g_1)$ and $g_1 = u_1 + (1 - u_1)(1 - T_1)$, [45]. In the case that $n = 2$ [49], we have $A_2 = G_0(g_1) - G_0(g_2)$, where $u_2 = G_1(g_2)/u_1$ and

$$\begin{aligned} g_2 =& u_1(u_2 + (1-u_2)(1-T_2)) \\ &+ (1-u_1)(1-T_1) \end{aligned} \tag{8.5}$$

Similarly, for $n = 3$ we have $A_3 = G_0(g_1)G_0(g_2) - G_0(g_3)$ with $u_3 = G_1(g_3)/(u_1 \cdot u_2)$ and

$$\begin{aligned} g_3 =& u_1(u_2(u_3 + (1-u_3)(1-T_3)) \\ &+ (1-u_2)(1-T_2)) + (1-u_1)(1-T_1) \end{aligned} \tag{8.6}$$

With these examples, it is hopefully clear how to write further generations of the competitive percolation process.

8.1.2 R_0

The R_0 value is defined as the number of new infections caused by an average infected individual. When $R_0 < 1$, the epidemic fails to infect a significant portion of the network; the GCC comprises $O(1)$ vertices. When $R_0 = 1$, the probability that an epidemic infects a macroscopic fraction of the network, $O(N)$, is non-zero. This point is also the critical point in bond percolation that marks the smallest value of T that can form a GCC. Within our model, there is an $R_{0,i}$ value and a critical transmissibility for each strain. This critical point is a function of both the network topology and the transmissibilities of the previous strains. If the transmissibility of a particular strain is below its critical threshold, it only infects $O(1)$ vertices before it burns out. Therefore, in the following analysis, we assume that the transmissibility of each strain is greater than its minimum threshold.

The critical point for the i-th percolation can be found by applying linear stability analysis around the fixed point $u_i = 1$. This is the point at which the fixed point in u_i bifurcates into two solutions and A_i becomes non-zero. Performing this analysis, we find the following condition

$$R_{0,i} = \prod_j u_j \left[\frac{\partial G_1(g_i)}{\partial g_i} \frac{\partial g_i}{\partial u_i} \right]^{-1}_{u_i=1} \tag{8.7}$$

where the derivatives are given by

$$\frac{\partial g_i}{\partial u_i} = T_i \prod_{j=1}^{i-1} u_j \tag{8.8}$$

When evaluated at $u_i = 1$, $G_1'(g_i)$ becomes $G_1'(g_{i-1})$ from Eqs 8.1 and 8.2. The critical transmissibility is found when $R_{0,i} = 1$ to be

$$T_{i,c} = \frac{1}{G_1'(g_{i-1})} \tag{8.9}$$

Thus, the minimum transmissibility required for each strain to create an epidemic is a function of the network topology and the transmissibilities of the preceding strains. Given that the coefficients of $G_1'(x)$ are non-negative and therefore monotonically increasing on the positive real line (within its radius of convergence), and that $g_j \in [0,1]$ (since it is a probability) and because $T_j, u_j \leq 1$, then it follows that $G_1'(g_i) \leq G_1'(g_{i-1})\ \forall i$. This indicates (from Eq 8.7) that $T_{i,c} \geq T_{i-1,c}$. In other words, the epidemic threshold of each strain increases with each generation. This is an intuitive result, since, as each strain passes through the network, vertices with higher degree are preferentially embedded into the GCC of that strain. Therefore, the RG is increasingly comprised of lower degree vertices as it fractures with each iteration of the percolation.

Following [26] we can also prove a stronger condition on the minimum bond occupation probability that a subsequent strain must have in order to exhibit an epidemic on the network. It happens that each generation of the disease must have an increasingly higher transmissibility than the last in order to infect $O(N)$ vertices in the RG. To see this, we note that $g_i(\boldsymbol{T}, \boldsymbol{u})$ is the probability that an edge fails to connect a vertex to the GCC of the i-th epidemic and that this probability can only decrease or stay constant as T_i increases;

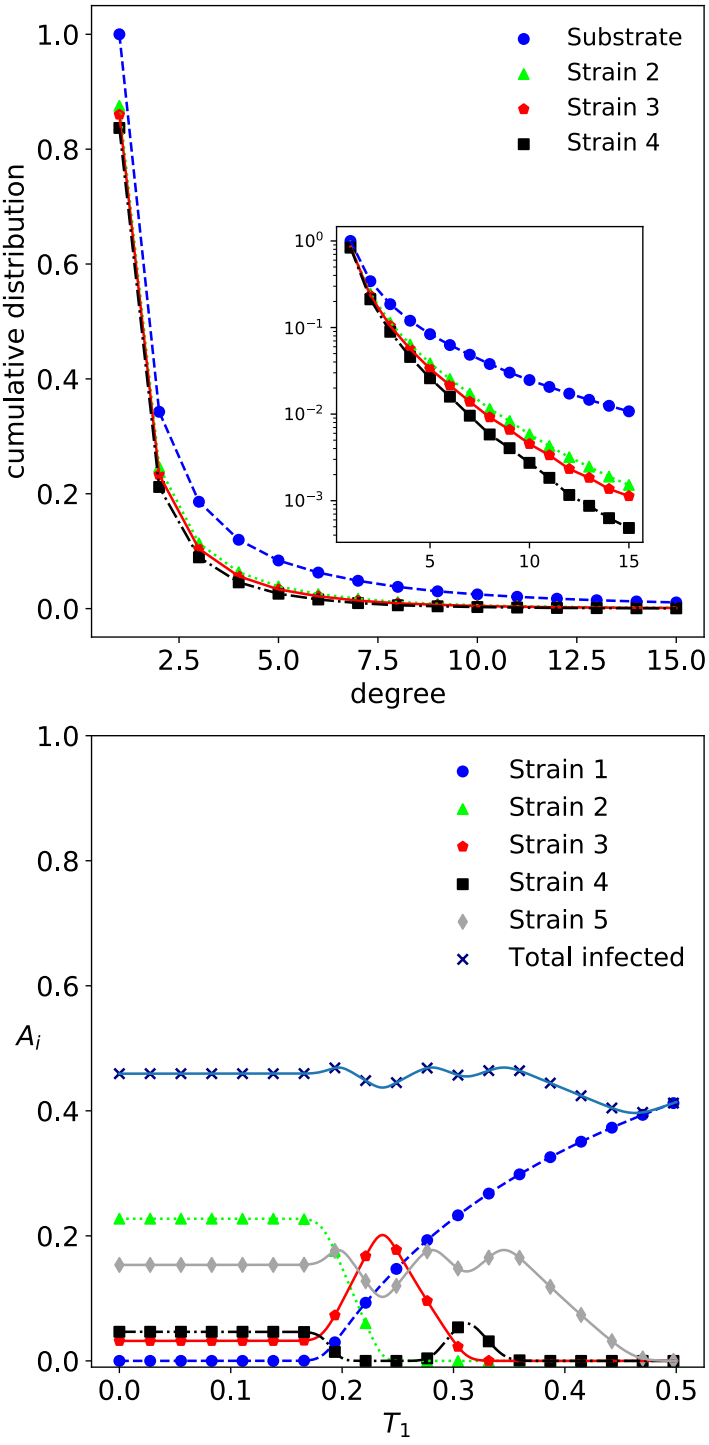


Figure 8.3: (top) The cumulative degree distribution of vertices in the substrate network and the GCCs of the RG following competitive bond percolation at $T_1 = 0.0$. (bottom) The outbreak sizes of the competitive branching process. The parameters are $(T_2, T_3, T_4, T_5) = (0.3, 0.5, 0.6, 1.0)$ on a scale-free network with power-law exponent $\alpha = 2$ and $\kappa = 20$. Scatter points are the average of 35 repeats of Monte Carlo simulations over $N = 30000$ vertex networks; whilst, solid lines are the theoretical results.

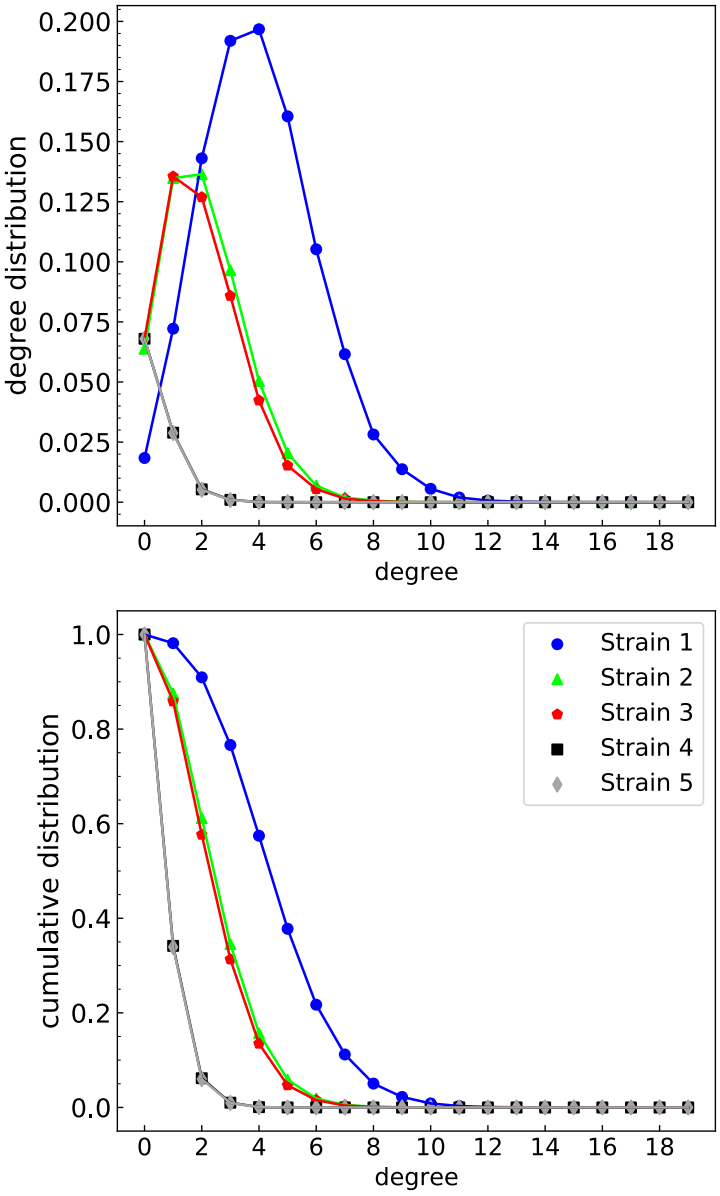


Figure 8.4: (top) The degree distribution of the RG created once each strain from Fig 8.2 has spread over the network at $T_1 = 0.0$; and, the corresponding cumulative degree distribution (bottom). Scatter points are the average of 50 repetitions of $N = 35000$ Erdős-Renyi networks with $\langle k \rangle = 4$. Curves are the theoretical results from the model. These plots show how the RG becomes increasingly fractured with each percolation.

this implies that $dT_i/dg_i \leq 0$. Inverting this quantity such that $T_i = T_i(\boldsymbol{u}, g_i, \boldsymbol{T}\backslash\{T_i\})$ where the notation $\boldsymbol{S}\backslash\{s\}$ excludes element s from set $\boldsymbol{S}$, and performing the derivative we have an expression that involves T_i and $G_1'(g_{i-1})$. This can then be isolated and it can be shown that $T_{i,c} \geq T_{i-1}\ \forall i \in [1,n]$. For example, the critical point of strain 2 is known [49, 26] to be greater than the transmissibility of strain 1, T_1. The critical point of strain 3 is given by

$T_{3,c} = 1/G_1'(g_2)$ from Eq 8.7 which, through the above prescription satisfies

$$T_{3,c} \geq \frac{g_1 - g_2}{1 - u_2} = T_2 \tag{8.10}$$

This logic can be applied to all adjacent strains to create an ordered set of critical transmissibilities $\{T_{1,c} \leq T_{2,c}, \ldots, T_{n,c}\}$. This indicates that transmissibility must evolve to increase in order for a given strain to create an epidemic in the presence of others.

8.1.3 Coexistence threshold

The coexistence threshold T_x [49, 26] for two pathogens marks an additional phase transition in the model. It is the largest value of T_1 that still allows the RG to retain sufficient connectivity to support its own GCC for future strains. For instance, when $T_1 > T_x$, the RG fails to be globally connected and strain 2 fails to infect $O(N)$ vertices even if $T_2 > T_{2,c}$. For our purpose we extend the definition of the coexistence threshold, $T_{i,x}$, in the context of n sequential strains to be the largest transmissibility of strain-i that allows the RG to support a GCC for future generations, assuming that they are sufficiently transmissible. Thus, $T_{i,x}$ is a function of the bond occupancy probabilities of all previous percolations, $T_{i,x} = T_{i,x}(\boldsymbol{u}, \boldsymbol{T} \backslash \{T_i\})$. As for $n = 2$ [49, 26], Eq 8.7 implicitly defines the coexistence threshold of the i-th strain and we find that $T_{i,x}$ is the value of T_i for which $G_1'(g_i) = 1$. For instance, for an Erdős-Renyi degree distribution this condition becomes

$$\frac{1}{\langle k \rangle} \ln \left[u_i \prod_{j=1}^{i-1} u_j \right] = g_i - 1 \tag{8.11}$$

from which we can solve for $T_{i,c}$ by inverting g_i. For $i = 1$ we have

$$T_{1,x} = \frac{\ln(u_1)}{\langle k \rangle} \frac{1}{(u_1 - 1)} \tag{8.12}$$

In Fig 8.2 we plot A_i for $n = 5$ against $T_1 \in [0,1]$ and $T_2 = 0.35$, $T_3 = 0.5$, $T_4 = 1.0$ and $T_5 = 1.0$ for a Erdős-Renyi random graph with mean degree $\langle k \rangle = 4$ and $N = 30000$ vertices. We observe excellent agreement between experimental bond percolation (scatter points) and the analytical results of Eq 8.4 (plotted lines). Below the epidemic threshold of the first strain, $T_1 < T_{1,c}$, strain 1 does not exhibit a GCC. Hence, the RG is large enough to enable the subsequent strains to form their own GCCs, each consuming more of the available space. With $T_4 = 1$, the last edges in the RG are occupied and, despite a supercritical T_5, we have $A_5 = 0$. Strain-4 is bimodal, exhibiting two turning points as a function of T_1. This is because, at the first turning point, the transmissibility of the previous strains is sufficient to form their own large GCCs in the RG; however, as T_1 increases, strains 2 and 3 fall below their critical thresholds, allowing strain-4 to consume the available sites into its own GCC. The outbreak size of strain-4 then falls to zero through a final turning point as the transmissibility of strain-1 is increased beyond the coexistence threshold. The inset figure shows the total fraction of the network that has become infected, A versus T_1. Against intuition, the largest fraction of the network that is occupied by (any) disease, is not constant. To see this, we understand that the early generations of the disease consume the high degree sites. As these become embedded within the GCC, those vertices they connect to can become isolated and thus, cannot be incorporated in subsequent GCCs

(see Fig 8.4). The effect of this is prominent at the onset of the GCC of strain 1 and leads to a local minimum in the total infecteds. Therefore, a disease of low transmissibility early on not only consumes vertices into its own GCC, but also reduces the accessible sites by removing the infection pathways. As the transmissibility of the initial pathogen increases, this effect is reduced and the total infected fraction of the network increases to $A \approx 0.9$. The global minimum at $T_1 \approx 0.45$ coincides with the coexistence threshold for strain 1; we observe the inability of subsequent strains to create their own epidemic. At this point, strain 1 is sufficiently transmissible to fracture the RG such that it can no longer support a GCC for the other generations of the disease. Beyond this point the total infected fraction follows the number of infected vertices of strain 1.

8.2 N-strain coinfection

In this section we define the i-th generation of a collaborative branching process as a bond percolation process occurring on the GCC that is created by the $i-1$ previous processes for $i = 1, \ldots, n$. We impose the strict requirement that only vertices in the GCC created by all of the previous generations are included at the i-th generation. This process has been studied previously by Newman and Ferrario when $n = 2$ [53] as well as for clustered and modular networks [31]. Within the context of the SIR isomorphism, this model studies the ability of the i-th disease to become an epidemic given that coinfection with all other $i-1$ strains is a prerequisite for infection with the current strain. The model is slightly more complex than the competitive percolation and is best understood in terms of epidemiology, so we will continue to use that setting for this section. If a vertex fails to become infected with a particular strain, then it cannot become infected with further generations of the disease; strains that have a low transmissibility significantly reduce the pool of available vertices for future outbreaks.

To describe the model we index the generations $i \in [1, n]$ as before. We choose a vertex at random from the GCC of the $(i-1)$-th percolation, prior to the i-th percolation. The neighbours of the vertex are either attached to the GCC or they belong to the residual graph with probability u_i. Of those neighbours attached to the GCC, a fraction have been *directly* attached by the particular chosen vertex, whilst the others have been attached by one of their other neighbours. In the context of epidemiology, we emphasize the particular subset of infected neighbours which the focal vertex directly infected from those that where infected by one of their other neighbours. A rich and complex epidemiological lineage, which we call the *infection history*, can be written for any neighbour given a particular focal vertex and generation index. For instance, the focal vertex could transmit strain-1 to a particular neighbour, which in turn, might be the vertex that transmits strain-2 to the focal vertex. It happens that the probability of passing each strain around is quite different depending on the precise infection history of each neighbour; the direction of infection transmission over a particular edge is important. Therefore, we must explicitly track all possible neighbour states and define a probability, u_{i_h}, that each possible edge fails to connect the focal vertex to the i-th GCC for a given history i_h. There are as many u_{i_h} values as there are independent maximally coinfected (states which have contracted all possible strains) states in the i-th generation, which we now examine in detail.

As each percolation unfolds over the network, the number of states that neighbouring vertices can occupy increases. Each generation branches the current number of maximally coinfected states by a factor of two (accounting for directly and indirectly infected neigh-

bours of the coinfected vertices in generation $i-1$). This means that the total number of neighbour states, η_i, in the i-th generation is

$$\eta_i = 1 + \sum_{j=1}^{i} 2^j \tag{8.13}$$

comprising all of the states in the previous generation (which did not contract the i-th strain) in addition to 2^{i-1} new states that are directly and indirectly infected, plus the uninfected branch. Thus, the set of all infection histories for for a given generation i, $\{h\}_i$, has cardinality 2^{i-1}; therefore, each generation requires 2^{i-1} new u_{i_h} values, with $i_h \in \{h\}_i$, $h = 1,\ldots,2^{i-1}$.

For instance, with reference to Fig 8.5, a vertex in the GCC of the first percolation has three neighbour-types represented by the three triangles: susceptible (unfilled triangle), directly (checked triangle) and indirectly (filled triangle) infected vertices. A vertex in the GCC of the second percolation has seven neighbour-types: susceptible (unfilled triangle), directly (checked triangle) and indirectly (filled triangle) infected with strain-1 but *not* infected by strain-2; and finally, coinfected states (squares and pentagons) via direct (checked) and indirect (filled) infection with strain 2 over both the direct (squares) and indirect (pentagons) branches of strain-1. The third generation has 15 potential vertex states comprising the seven states from the previous generation and twice the number of coinfected states 2×4 and so on.

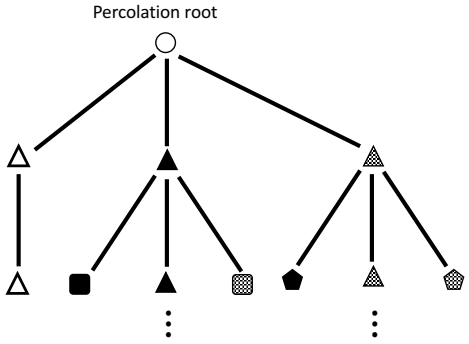


Figure 8.5: The possible neighbour vertex states surrounding a particular focal vertex for the first two generations of the collaborative branching process. Each level of the tree represents sequential generations of percolation, starting from an un-percolated root, to which all vertices belong. Unfilled vertices represent states that do not belong to the GCC, solid vertices represent states that are externally infected whilst checked vertices represent states that have been directly infected by the focal vertex. For instance, the filled square of the third generation has been externally infected with disease 1 and disease 2; whilst the checked pentagon has been directly infected by the focal for both strains. All states from the $(i-1)$-th percolation are brought forward into the i-th level, representing failed infection by generation i.

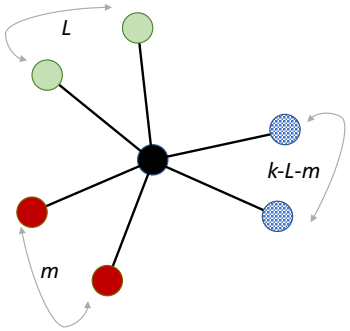


Figure 8.6: A graphical representation of a degree k focal vertex (center) with l uninfected, m externally infected and $k-l-m$ directly infected neighbours.

8.2.1 Outbreak size

The aim of this section is to define a prescription to obtain the outbreak size of the i-th epidemic. Each generation requires 2^{i-1} unique u_{i_h} values to be written; each accounting for a particular infection history that a neighbouring vertex could have. Each u_{i_h} value will then be generated by a self-consistent expression in a similar way to those of the competitive model as

$$u_{i_h} = \frac{G_1(P_{i_h})}{Q_{i_h}} \tag{8.14}$$

where P_{i_h} is the probability of not obtaining strain i for infection history i_h and Q_{i_h} is the prior probability that the neighbour has infection history i_h. Note, P_{i_h} is analogous to g_i from the competitive model. Thus, it remains to calculate both the prior probabilities and P_{i_h}. It happens, for a given generation i, that the probabilities P_{i_h} can be factored; thus, each P_{i_h} expression comprises two parts: a multiplying common factor, C_i; and, the unique part of the probability of each specific infection history, H_{i_h} such that

$$P_{i_h} = C_i H_{i_h} \tag{8.15}$$

Firstly, we calculate C_i, which is simply all of the common terms belonging to each P_{i_h}, $\forall i_h \in \{h\}_i$. Consider each branch point of the collaborative process from the perspective of an infected vertex as we progress from generation $j-1$ to j. Neighbours either do not contract strain-j; or they do, from either the focal vertex or one of their other neighbours; there are always three branches from a given state (see Fig 8.6). Let $f_j(u_{j_h}, v, w)$ be a function that encapsulates the three possible neighbour scenarios and let u_{j_h} be the probability that a neighbour is uninfected by any of its other neighbours by the j-th strain at this branch point, given that its infection history is j_h. We have

$$f_j(u_{j_h}, v, w) = \sum_{l=0}^{k} \binom{k}{l} [u_{j_h}(1-T_j)]^l \sum_{m_{j_h}=0}^{k-l} \binom{k-l}{m_{j_h}} [(1-u_{j_h})v]^{m_{j_h}} [u_{j_h} T_j w]^{k-l-m_{j_h}} \tag{8.16}$$

We have indexed the variable m with the infection history for later convenience. Despite the complicated form of this expression it is straightforward to construct each term by considering the probabilities associated with each neighbouring state. In detail: $u_{j_h}(1-T_j)$

is the probability that a neighbour was uninfected by its other neighbours given history h and that the focal vertex did not transmit strain j; $1-u_{j_h}$ is the probability that a neighbour was already infected and $u_{j_h}T_j$ is the probability that a neighbour was directly infected by the focal vertex. The arguments v and w are placeholders that allow the further subdivision of the number of neighbours in a given infected state following the next generation.

We construct the common factor C_i by first constructing a related factor, $\bar{C}_i$, composing this branch-point logic with itself i times and terminating the composition with $v = w = 1$ at the deepest levels (i.e. the leaves of the branching process) such that

$$\bar{C}_i = f_1(f_2(\ldots f_i())) \tag{8.17}$$

The function $\bar{C}_i$ has 2^{i-1} arguments. The values of j_h in u_{j_h} are given by the particular elements of $\{h\}_j$ $j = 1,\ldots,i-1$. The common factor for the percolation root (prior to any diseases) is unity, $\bar{C}_0 = 1$; since all strains belong to the same state. Following strain-1 $\bar{C}_1$ is given by

$$\begin{aligned}\bar{C}_1(1,1) &= f_1(u_{1_1},1,1)\\ &= u_{1_1}(1-T_1)+1-u_{1_1}+u_{1_1}T_1\end{aligned} \tag{8.18}$$

for $i = 2$ we have

$$\begin{aligned}\bar{C}_2(1,1,1,1) &= f_1(u_{1_1},f_2(u_{2_1},1,1),f_2(u_{2_2},1,1))\\ &= u_{1_1}(1-T_1)+(1-u_{1_1})(u_{2_1}(1-T_2)+1-u_{2_1}\\ &\quad +u_{2_1}T_2)+u_{1_1}T_1(u_{2_2}(1-T_2)+1-u_{2_2}+u_{2_2}T_2)\end{aligned} \tag{8.19}$$

similarly, for $i = 3$ we have

$$\bar{C}_3(\mathbf{1}) = f_1(u_1,f_2(u_{2_1},f_3(u_{3_1},1,1),f_3(u_{3_2},1,1)),f_2(u_{2_2},f_3(u_{3_3},1,1),f_3(u_{3_4},1,1))) \tag{8.20}$$

and so on. An interesting observation is that this expression is always unity and that there are always as many terminating 1s as there are unique infection histories required for the next generation. We must define another related probability, $\bar{f}_i$, as

$$\bar{f}_j(u_j,v,w) = \sum_{l=0}^{k}\binom{k}{l}[u_j(1-T_j)]^l\sum_{m=0}^{k-l}\binom{k-l}{m}[(1-u_j)(1-T_j)v]^m[u_jT_jw]^{k-l-m} \tag{8.21}$$

which is the probability that none of the externally infected neighbours transmitted their infection to the focal vertex. Given these two functions, we can now build the common terms in the probability that the focal vertex does not contract the i-th strain as

$$C_i = \bar{C}_{i-1}(\bar{\boldsymbol{f}}_{\boldsymbol{i}}) \tag{8.22}$$

Thus, C_i contains the common terms in the probability that describes the neighbouring states prior to strain i, along with the probability that each of those states then fails to

transmit strain i itself. For example, the first few values of C_i are

$$C_1 = \bar{f}_1(u_1, 1, 1) \tag{8.23a}$$
$$C_2 = \bar{C}_1(\bar{f}(u_{2_1}, 1, 1), \bar{f}(u_{2_2}, 1, 1)) \tag{8.23b}$$
$$C_3 = \bar{C}_2(\bar{f}_3(u_{3_1}, 1, 1), \bar{f}_3(u_{3_2}, 1, 1), \bar{f}_3(u_{3_3}, 1, 1), \bar{f}_3(u_{3_4}, 1, 1)) \tag{8.23c}$$

With a clear prescription to derive C_i for each generation, we must now calculate the

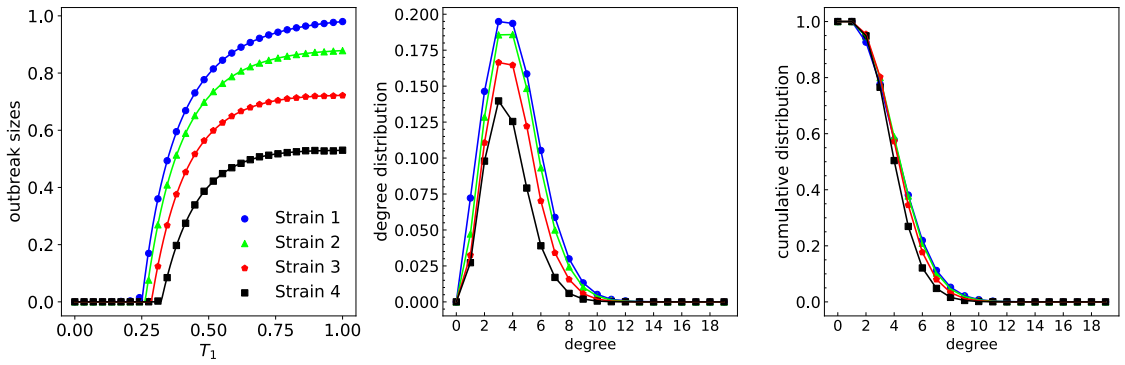


Figure 8.7: Four generations of the cooperative branching process with $(T_2, T_3, T_4) = (0.6, 0.5, 0.45)$ and on an Erdős-Renyi network with mean degree $\langle k \rangle = 4$. Scatter points are the average of 35 repeats of Monte Carlo simulations over $N = 30000$ vertex networks; whilst, solid lines are the theoretical results. Subplot (a) shows the outbreak sizes; (b) shows the degree distribution at $T_1 = 1$; (c) is the cumulative probability that a vertex has degree larger than k at $T_1 = 1$.

probability associated with each infection history H_{i_h} in order to finalise the expressions for P_{i_h}, which in turn we require in order to write self-consistent expressions for each u_{i_h} value in Eq 8.14. To do this, we consider each pathway from the percolation root to the leaves of the tree created by the collaborative branching process (see Fig 8.5). If, at a particular branching point, we progress *via* direct infection, we require that the focal vertex was the vertex that transmitted infection to the neighbour. For this to occur we require that the other neighbours other than the focal vertex failed to transmit their infection. This occurs with probability $(1-T_j)^{m_r}$, with reference to Eq 8.16. Similarly, the probability that the neighbour *was* infected by a vertex other than the focal vertex is $1-(1-T_j)^{m_r}$. We now see the utility of subscripting m in Eq 8.16 as it allows us to track each particular set of externally infected neighbours over the arguments of C_i. For instance, there are two infection histories at the start of the second process, strain-1 infected vertices have either been externally infected, 2_1, or directly infected by the focal vertex, 2_2, such that $\{h\}_2 = \{2_1, 2_2\}$. The number of externally 1-infected neighbours is given by m_{1_1} and so we have

$$H_{2_1} = [1-(1-T_1)^{m_{1_1}}] \tag{8.24}$$
$$H_{2_2} = (1-T_1)^{m_{1_1}} \tag{8.25}$$

Similarly, a vertex can obtain strain-3 from one of four different neighbour states: externally-1 and externally-2 infected (3_1); externally-1 and directly-2 infected (3_2); directly-1 and

externally-2 infected (3_3), or finally, directly-1 and directly-2 infected (3_4). Thus, there are four infection histories to generate with $\{h\}_4 = \{3_1, 3_2, 3_3, 3_4\}$. We then write

$$H_{3_1} = H_{2_1}[1-(1-T_2)^{m_{2_1}+m_{2_2}}] \tag{8.26a}$$
$$H_{3_2} = H_{2_1}(1-T_2)^{m_{2_1}+m_{2_2}} \tag{8.26b}$$
$$H_{3_3} = H_{2_2}[1-(1-T_2)^{m_{2_1}+m_{2_2}}] \tag{8.26c}$$
$$H_{3_4} = H_{2_2}(1-T_2)^{m_{2_1}+m_{2_2}} \tag{8.26d}$$

Each history is constructed from the necessary probabilities to create each scenario. For the next generation, each of these unique infection histories are branched into two to give eight potential sources of strain-4.

With these examples, we now have a prescription to write P_{i_h} for each potential infection neighbouring state. The last component we require in order to calculate the associated u_{i_h} values in Eq 8.14 is the prior probabilities that the neighbour was indeed in that particular state following all of the previous strains, but prior to the i-th strain itself. It happens that this probability follows a recipe that is simple to compute for each term. We define the following rule: if a neighbour is externally infected at the j-th strain, we multiply the prior probability by $(1-u_{j_h})$; otherwise, for direct infection, we multiply by u_{j_h} instead. The logic behind this rule is simple: to be directly infected by the focal vertex, a neighbour must not be infected by their other neighbours; the focal vertex *must* be the successful infection pathway. Therefore, for the first branch point, the prior probabilities that the neighbour was externally and directly infected, respectively, are given by

$$Q_{2_1} = (1-u_{1_1}) \tag{8.27}$$
$$Q_{2_2} = u_{1_1} \tag{8.28}$$

Similarly following the second strain the prior probabilities for the four infection histories are

$$Q_{3_1} = (1-u_{1_1})(1-u_{2_1}) \tag{8.29a}$$
$$Q_{3_2} = (1-u_{1_1})u_{2_1} \tag{8.29b}$$
$$Q_{3_3} = u_{1_1}(1-u_{2_2}) \tag{8.29c}$$
$$Q_{3_4} = u_{1_1}u_{2_2} \tag{8.29d}$$
$$\tag{8.29e}$$

With this last component we can now construct the self-consistent expressions required to compute the u_{i_h} values in Eq 8.14. At this point, a useful check to ensure that the derived components (priors, histories and base term) are correct is to set $\bar{\boldsymbol{f}}_i = \mathbf{1}$. When evaluated at unity, the u_{i_h} values should be equal to one.

Next, we require the outbreak size of the i-th strain, A_i. It happens that this expression is very simple to construct once we have performed the above work; we simply have to take the expression for $u_{i,h}$ for the maximally indirect, maximally coinfected infection history i_1 (i.e. the expression for the history where every vertex-state was externally infected), remove the prior denominator and replace the $G_1(z)$ generating function with a $G_0(z)$ generating function. We then subtract this value from the previous outbreak size such that

$$A_i = A_{i-1} - G_0(P_{i_1}) \tag{8.30}$$

Since $G_0(P_{i_1})$ can never be negative, we observe that the outbreak size of each generation can never exceed the size of the previous one. We detail the expressions for the first few generations in section 8.3 and show the percolation results for the first four generations spreading over an Erdős-Renyi network in Fig 8.7. In Fig 8.7 (a) we plot the outbreak sizes of each strain; each exhibits a smaller size and a larger percolation threshold with increasing strain index. In Fig 8.7 (b) the degree distribution of each of the GCC substructures is plotted. The average degree is reduced and the height and variance of each distribution is increasingly reduced and shifted to the left. In Fig 8.7 (c) the cumulative probability that the degree of a vertex is larger than k is shown for each strain. These results indicate that eventually the spreading of cooperative processes on Erdős-Renyi graphs will be limited by the fractured topology of the substrate network available to each strain in addition to the transmissibility of the disease.

The complete prescription for solving for the outbreak size of the n-th generation of the cooperative branching process is to hierarchically solve the coupled linear system of equations for each u_{i_j} given by

$$u_{i_j} = u_{i_j}(u_{1_1}, u_{2_1}, \ldots, u_{i_{2^{i-1}}}; T_1, \ldots, T_i) \tag{8.31}$$

for infection histories $j = 1, \ldots, 2^{i-1}$, and generations $i = 1, \ldots, n$ and functional form given by Eq 8.14. More detail on the structure of these expressions in terms of a perfect binary tree is treated in section 8.3 as well as an examination of their solutions.

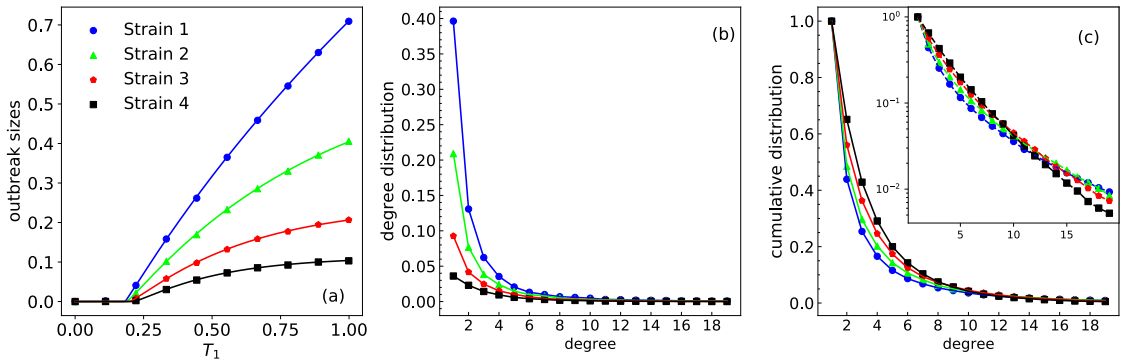


Figure 8.8: Four generations of the cooperative branching process with $(T_2, T_3, T_4) = (0.6, 0.5, 0.45)$ and on a scale-free network with power-law exponent $\alpha = 2$ and $\kappa = 20$. Scatter points are the average of 35 repeats of Monte Carlo simulations over $N = 30000$ vertex networks; whilst, solid lines are the theoretical results. Subplot (a) shows the outbreak sizes, (b) shows the degree distribution at $T_1 = 1$; whilst, (c) is the cumulative probability that a vertex has degree larger than k in each GCC at $T_1 = 1$; the inset shows the same data on a logarithmic scale.

8.3 Outbreak sizes for collaborative branching processes

In this section we describe how to use the prescription to obtain the outbreak sizes of the first few generations of the collaborative branching process. We also examine the graphical solution for the resulting non-linear system of equations for each strain. This allows us to

examine the relative contribution of each infection mode to the overall outbreak size of the epidemic.

8.3.1 Strain 2

For strain 2 there are two possible infection histories that a neighbour might have: either the focal vertex infected it directly or it was externally infected. The probability that the focal vertex doesn't get strain 2 from each of these neighbour states is u_{2_1} and u_{2_2}, respectively. The common factor is given by

$$C_2 = \bar{C}_1(\bar{f}(u_{2_1},1,1),\bar{f}(u_{2_2},1,1)) \tag{8.32}$$

which is simply

$$C_2 = \sum_{l=0}^{k}\binom{k}{l}[u_{1_1}(1-T_1)]^l \sum_{m_{1_1}=0}^{k-l}\binom{k-l}{m_{1_1}}[(1-u_{1_1})\bar{f}(u_{2_1})]^{m_{1_1}}[u_{1_1}T_1\bar{f}(u_{2_2})]^{k-l-m_{1_1}} \tag{8.33}$$

where we have dropped the 1s in the function arguments of the $\bar{f}$ functions. Following the prescription, the history of u_{2_1} is $H_{2_1} = [1-(1-T_1)^{m_{1_1}}]$ whilst the history of u_{2_2} is $H_{2_2} = (1-T_1)^{m_{1_1}}$. Since the u_{2_1} history branches from the $1-u_{1_1}$ compartment, the prior probability is simply $Q_{2_1} = 1-u_{1_1}$; whilst $Q_{2_2} = u_{1_1}$. Thus, we have

$$u_{2_1} = \frac{1}{Q_{2_1}}\sum_{k=0}^{\infty} q_k C_2 H_{2_1} \tag{8.34}$$

$$u_{2_2} = \frac{1}{Q_{2_2}}\sum_{k=0}^{\infty} q_k C_2 H_{2_2} \tag{8.35}$$

We exhibit the graphical solution for these coupled equations in Fig 8.9.

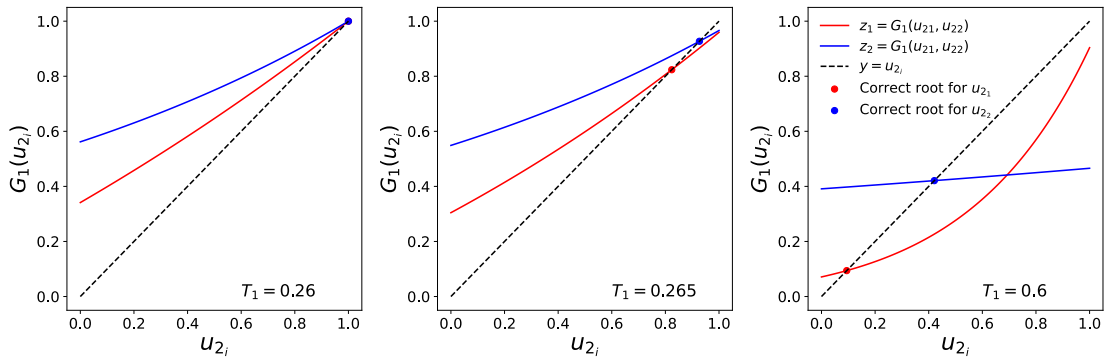


Figure 8.9: The graphical solution of the generating functions for strain 2 at three different T_1 values with $T_2 = 0.6$. Plotted are $y = u_{2_i}$ for $i = 1,2$ against u_{2_i} as well as the value of the generating functions $z_1 = G_1(C_2H_{2_1})/(1-u_{1_1})$ and $z_2 = G_1(C_2H_{2_2})/u_{1_1}$. Each z_i varies u_{2_i} whilst the other value is held fixed at the correct root. The intersection of $y = u_{2_i}$ and z_i corresponds to the root, which is also marked with a scatter point. We notice that the trivial root of $u_{2_1} = u_{2_2} = 1$ is no longer shown as the system moves away from the critical point when the GCC first forms. This is also graphical motivation for finding the critical point from a Taylor series around the trivial root. We also notice the increase (loss) of convexity in u_{2_1} (u_{2_2}) as we increase T_1. This indicates the increasing (decreasing) importance of the u_{2_1} (u_{2_2}) branch to the formation of the GCC at larger transmissibilities.

8.3.2 Strain 4

We have

$$\begin{aligned} C_4 = f_1(u_1, f_2(u_{2_1}, f_3(u_{3_1}, \bar{f}_4(u_{4_1}), \bar{f}_4(u_{4_2})), f_3(u_{3_2}, \bar{f}_4(u_{4_3}), \bar{f}_4(u_{4_4}))), \\ \times f_2(u_{2_2}, f_3(u_{3_3}, \bar{f}_4(u_{4_5}), \bar{f}_4(u_{4_6})), f_3(u_{3_4}, \bar{f}_4(u_{4_7}), \bar{f}_4(u_{4_8})))) \end{aligned} \tag{8.36}$$

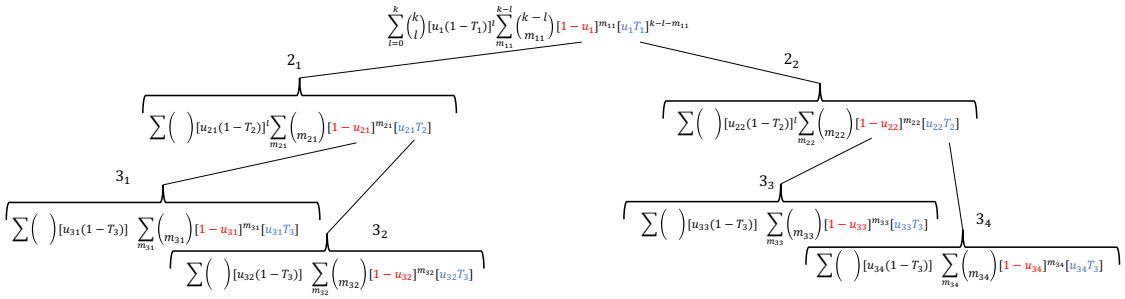


Figure 8.10: A visualisation of the perfect binary tree of coinfected neighbours which a focal vertex could be surrounded by and the equations that generate the base probability $\bar{C}_3(\mathbf{1})$. In this notation, the left child always represents external infection whilst the right child represents direct infection. Thus, the leaves of the tree represent all possible infection histories for the 4th strain.

$$u_{4_1} = \frac{1}{(1-u_{1_1})(1-u_{2_1})(1-u_{3_1})} \sum_{k=0}^{\infty} q_k C_4 H_{3_1} [1-(1-T_3)^{m_{3_1}+m_{3_2}+m_{3_3}+m_{3_4}}] \tag{8.37a}$$

$$u_{4_2} = \frac{1}{(1-u_{1_1})(1-u_{2_1})u_{3_1}} \sum_{k=0}^{\infty} q_k C_4 H_{3_1} (1-T_3)^{m_{3_1}+m_{3_2}+m_{3_3}+m_{3_4}} \tag{8.37b}$$

$$u_{4_3} = \frac{1}{(1-u_{1_1})u_{2_1}(1-u_{3_2})} \sum_{k=0}^{\infty} q_k C_4 H_{3_2} [1-(1-T_3)^{m_{3_1}+m_{3_2}+m_{3_3}+m_{3_4}}] \tag{8.37c}$$

$$u_{4_4} = \frac{1}{(1-u_{1_1})u_{2_1}u_{3_2}} \sum_{k=0}^{\infty} q_k C_4 H_{3_2} (1-T_3)^{m_{3_1}+m_{3_2}+m_{3_3}+m_{3_4}} \tag{8.37d}$$

$$u_{4_5} = \frac{1}{u_{1_1}(1-u_{2_2})(1-u_{3_3})} \sum_{k=0}^{\infty} q_k C_4 H_{3_3} [1-(1-T_3)^{m_{3_1}+m_{3_2}+m_{3_3}+m_{3_4}}] \tag{8.37e}$$

$$u_{4_6} = \frac{1}{u_{1_1}(1-u_{2_2})u_{3_3}} \sum_{k=0}^{\infty} q_k C_4 H_{3_3} (1-T_3)^{m_{3_1}+m_{3_2}+m_{3_3}+m_{3_4}} \tag{8.37f}$$

$$u_{4_7} = \frac{1}{u_{1_1}u_{2_2}(1-u_{3_4})} \sum_{k=0}^{\infty} q_k C_4 H_{3_4} [1-(1-T_3)^{m_{3_1}+m_{3_2}+m_{3_3}+m_{3_4}}] \tag{8.37g}$$

$$u_{4_8} = \frac{1}{u_{1_1}u_{2_2}u_{3_4}} \sum_{k=0}^{\infty} q_k C_4 H_{3_4} (1-T_3)^{m_{3_1}+m_{3_2}+m_{3_3}+m_{3_4}} \tag{8.37h}$$

$$\tag{8.37i}$$

We show the solution to these expressions graphically in Fig 8.11 around the critical point and for large T values. The outbreak size is then given by

$$A_4 = A_3 - \sum_{k=0}^{\infty} p_k C_4 H_{3_1} [1-(1-T_3)^{m_{3_1}+m_{3_2}+m_{3_3}+m_{3_4}}] \tag{8.38}$$

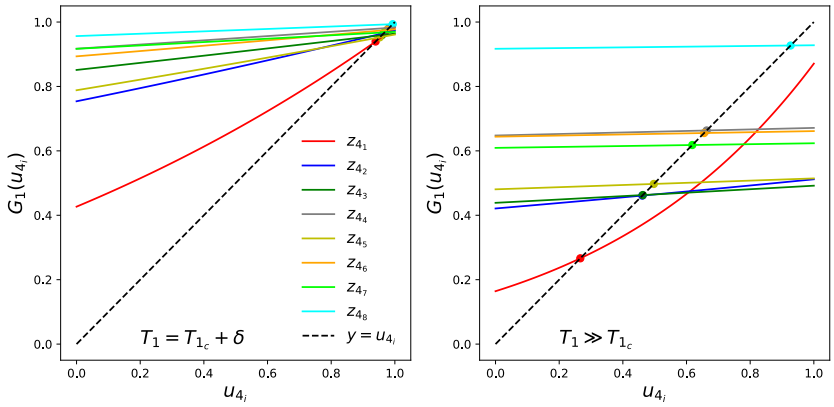


Figure 8.11: The graphical solution of the generating functions for strain 4 around the critical point of strain 1 (left) with $T_1 = 0.325, T_2 = 0.6, T_3 = 0.5, T_4 = 0.45$ and away from the critical point at $T_1 = T_2 = 0.6, T_3 = 0.5, T_4 = 0.45$ (right). Plotted are $y = u_{4_i}$ for $i = 1, \ldots, 8$ against u_{4_i} as well as the value of the generating functions $z_i = G_1(C_2 H_{4_i})/Q_{4_i}$. Each z_i varies u_{4_i} whilst the other value held fixed at the correct root. The intersection of $y = u_{4_i}$ and z_i corresponds to the root, which is also marked with a scatter point from a non-linear solve. We notice that u_{4_1} varies convexly over almost the entire unit interval whilst the gradient of the other generating functions is increasingly flat in this region of the parameter space. This indicates that the contribution of these values is less important than that u_{4_1}. Thus, we can expect that this infection history is the dominant term in the non-linear system that describes strain 4.

8.3.3 Strain 5

Following the recipe, the outbreak size of strain 5 is calculated as follows.

$$\begin{aligned} C_5 = f_1(u_1, f_2(u_{2_1}, f_3(u_{3_1}, f_4(u_{4_1}, \bar{f}_5(u_{5_1}), \bar{f}_5(u_{5_2})), f_4(u_{4_2}, \bar{f}_5(u_{5_3}), \bar{f}_5(u_{5_4}))), \\ \times f_3(u_{3_2}, f_4(u_{4_3}, \bar{f}_5(u_{5_5}), \bar{f}_5(u_{5_6})), f_4(u_{4_4}, \bar{f}_5(u_{5_7}), \bar{f}_5(u_{5_8})))), \\ \times f_2(u_{2_2}, f_3(u_{3_3}, f_4(u_{4_5}, \bar{f}_5(u_{5_9}), \bar{f}_5(u_{5_{10}})), f_4(u_{4_6}, \bar{f}_5(u_{5_{11}}), \bar{f}_5(u_{5_{12}}))), \\ \times f_3(u_{3_4}, f_4(u_{4_7}, \bar{f}_5(u_{5_{13}}), \bar{f}_5(u_{5_{14}})), f_4(u_{4_8}, \bar{f}_5(u_{5_{15}}), \bar{f}_5(u_{5_{16}}))))) \end{aligned} \tag{8.39}$$

$$u_{5_1} = \frac{1}{(1-u_{1_1})(1-u_{2_1})(1-u_{3_1})(1-u_{4_1})} \sum_{k=0}^{\infty} q_k C_5 H_{4_1} [1-(1-T_4)^M] \tag{8.40a}$$

$$u_{5_2} = \frac{1}{(1-u_{1_1})(1-u_{2_1})(1-u_{3_1})u_{4_1}} \sum_{k=0}^{\infty} q_k C_5 H_{4_1} (1-T_4)^M \tag{8.40b}$$

$$u_{5_3} = \frac{1}{(1-u_{1_1})(1-u_{2_1})u_{3_1}(1-u_{4_2})} \sum_{k=0}^{\infty} q_k C_5 H_{4_2} [1-(1-T_4)^M] \tag{8.40c}$$

$$u_{5_4} = \frac{1}{(1-u_{1_1})(1-u_{2_1})u_{3_1}u_{4_2}} \sum_{k=0}^{\infty} q_k C_5 H_{4_2} (1-T_4)^M \tag{8.40d}$$

$$u_{5_5} = \frac{1}{(1-u_{1_1})u_{2_1}(1-u_{3_2})(1-u_{4_3})} \sum_{k=0}^{\infty} q_k C_5 H_{4_3} [1-(1-T_4)^M] \tag{8.40e}$$

$$u_{5_6} = \frac{1}{(1-u_{1_1})u_{2_1}(1-u_{3_2})u_{4_3}} \sum_{k=0}^{\infty} q_k C_5 H_{4_3} (1-T_4)^M \tag{8.40f}$$

$$u_{5_7} = \frac{1}{(1-u_{1_1})u_{2_1}u_{3_2}(1-u_{4_4})} \sum_{k=0}^{\infty} q_k C_5 H_{4_4} [1-(1-T_4)^M] \tag{8.40g}$$

$$u_{5_8} = \frac{1}{(1-u_{1_1})u_{2_1}u_{3_2}u_{4_4}} \sum_{k=0}^{\infty} q_k C_5 H_{4_4} (1-T_4)^M \tag{8.40h}$$

$$u_{5_9} = \frac{1}{u_{1_1}(1-u_{2_2})(1-u_{3_3})(1-u_{4_5})} \sum_{k=0}^{\infty} q_k C_5 H_{4_5} [1-(1-T_4)^M] \tag{8.40i}$$

$$u_{5_{10}} = \frac{1}{u_{1_1}(1-u_{2_2})(1-u_{3_3})u_{4_5}} \sum_{k=0}^{\infty} q_k C_5 H_{4_5} (1-T_4)^M \tag{8.40j}$$

$$u_{5_{11}} = \frac{1}{u_{1_1}(1-u_{2_2})u_{3_3}(1-u_{4_6})} \sum_{k=0}^{\infty} q_k C_5 H_{4_6} [1-(1-T_4)^M] \tag{8.40k}$$

$$u_{5_{12}} = \frac{1}{u_{1_1}(1-u_{2_2})u_{3_3}u_{4_6}} \sum_{k=0}^{\infty} q_k C_5 H_{4_6} (1-T_4)^M \tag{8.40l}$$

$$u_{5_{13}} = \frac{1}{u_{1_1}u_{2_2}(1-u_{3_4})(1-u_{4_7})} \sum_{k=0}^{\infty} q_k C_5 H_{4_7} [1-(1-T_4)^M] \tag{8.40m}$$

$$u_{5_{14}} = \frac{1}{u_{1_1}u_{2_2}(1-u_{3_4})u_{4_7}} \sum_{k=0}^{\infty} q_k C_5 H_{4_7} (1-T_4)^M \tag{8.40n}$$

$$u_{5_{15}} = \frac{1}{u_{1_1}u_{2_2}u_{3_4}(1-u_{4_8})} \sum_{k=0}^{\infty} q_k C_5 H_{4_8} [1-(1-T_4)^M] \tag{8.40o}$$

$$u_{5_{16}} = \frac{1}{u_{1_1}u_{2_2}u_{3_4}u_{4_8}} \sum_{k=0}^{\infty} q_k C_5 H_{4_8} (1-T_4)^M \tag{8.40p}$$

$$\tag{8.40q}$$

with $M = \sum_{j=1}^{8} m_{4_j}$. The outbreak size is then given by

$$A_5 = A_4 - \sum_{k=0}^{\infty} p_k C_5 H_{4_1} [1-(1-T_4)^M] \tag{8.41}$$

8.3.4 R_0

In this section we examine the critical points of the cooperative model, generalising the result of [53] for $n > 2$. As with the competitive percolation, there is an $R_{0,i}$ value for each generation or strain. If at any point the outbreak size of a generation is subcritical, then there can be no subsequent outbreaks as coinfection is a strong condition on the proliferation of future strains. However, assuming that the previous $i-1$ strains did indeed cause an $O(N)$ outbreak, then there is some point $T_{i,c}$ at which the i-th strain can also lead to a finite sized propagation if its transmissibility exceeds this value.

The critical point for the i-th percolation can be found by applying linear stability analysis around the fixed point $\{u_i^h\} = 1$, which is the trivial root of the system of equations for each generation (see section 8.3). The critical point of the first strain is identical to the results from the competitive branching process; however, it is prudent to review this result. Given that $u_{1_1} = 1$ at the critical point, we perform a Taylor expansion about $\varepsilon_{1_1} = 1 - u_{1_1}$ using Eq 8.14 and truncate it to 1st order to obtain

$$\varepsilon_{1_1} \approx 1 - G_1(f_1)\big|_{u_{1_1}=1} + G_1'(f_1)f_1'\big|_{u_{1_1}=1}\varepsilon_{1_1} \tag{8.42}$$

Rearranging this result, with $G_1(1) = 1$ and $f_1' = T_1$, we obtain $T_{1,c} = 1/G_1'(1)$ in accordance with Eq 8.9 at $i = 1$. For the second strain, we now have two variables to consider depending on the unique infection history of the neighbouring vertex. The critical point occurs when both u_{2_1} and u_{2_2} are unity (see section 8.3 for a graphical motivation of this) and we again perform a 1st order Taylor expansion about small parameter $\varepsilon_{2_j} = 1 - u_{2_j}$ to obtain the following coupled system

$$\begin{aligned}
\varepsilon_{2_1} &\approx \varepsilon_{2_1}\frac{\partial F_{2_1}}{\partial u_{2_1}} + \varepsilon_{2_2}\frac{\partial F_{2_1}}{\partial u_{2_2}} \\
\varepsilon_{2_2} &\approx \varepsilon_{2_1}\frac{\partial F_{2_2}}{\partial u_{2_1}} + \varepsilon_{2_2}\frac{\partial F_{2_2}}{\partial u_{2_2}}
\end{aligned} \tag{8.43}$$

where we have set the functional form of u_{i_j} in Eq 8.14 to $u_{i_j} = F_{i_j}$ and evaluate the derivatives at the fixed point $u_{2_1} = u_{2_2} = 1$. The derivatives are

$$\frac{\partial F_{2_1}}{\partial u_{2_1}} = T_2 - G_1'(1 - T_1 + u_{1_1}T_1)(1 - T_1)T_2 \tag{8.44}$$

$$\frac{\partial F_{2_1}}{\partial u_{2_2}} = \frac{u_{1_1}T_1}{1 - u_{1_1}}\left[1 - G_1'(1 - T_1 + u_{1_1}T_1)\right]T_2 \tag{8.45}$$

and

$$\frac{\partial F_{2_2}}{\partial u_{2_1}} = \frac{G_1'(1 - T_1 + u_{1_1}T_1)(1 - u_{1_1})(1 - T_1)}{u_{1_1}}T_2 \tag{8.46}$$

$$\frac{\partial F_{2_2}}{\partial u_{2_2}} = G_1'(1 - T_1 + u_{1_1}T_1)T_1T_2 \tag{8.47}$$

Thus, we have the following linear system

$$\boldsymbol{J}\begin{pmatrix} u_{2_1} \\ u_{2_2} \end{pmatrix} = \frac{1}{T_{2,c}}\begin{pmatrix} u_{2_1} \\ u_{2_2} \end{pmatrix} \tag{8.48}$$

where $\boldsymbol{J}$ is a Jacobian matrix with eigenvalue $1/T_{2,c}$. Following [53] the system has two eigenvalues and by examining the $T_1 \to 1$ limit, the correct eigenvalue is

$$T_{2,c} = \frac{2}{\tau + \sqrt{\tau^2 - 4\Delta}} \tag{8.49}$$

where τ is the trace of $\boldsymbol{J}$ and Δ is the determinant.

In the general case we have the following linear system

$$\varepsilon_{i_j} \approx \sum_{k=1}^{2^{i-1}} \frac{\partial F_{i_j}}{\partial u_{i_k}} \varepsilon_{i_k} \tag{8.50}$$

with $i \in [1,n]$ and $j \in [1,2^{i-1}]$. The derivatives are given by

$$\frac{\partial F_{i_j}}{\partial u_{i_k}} = \frac{G_1'(P_{i_j})}{Q(i_j)} \frac{\partial \bar{C}_{i-1} H_{i_j}}{\partial \bar{f}_i} \frac{\partial \bar{f}_i}{\partial u_{i_k}}\bigg|_{u_{i_k}=1} \tag{8.51}$$

The derivative of the final $\bar{f}_i$ term is always $\partial \bar{f}_i = T_i$ meaning that we have a leading factor of T_i multiplying all terms. Thus, we can create the following linear system by simple re-arrangement

$$\boldsymbol{J}\vec{u} = \frac{1}{T_{i,c}}\vec{u} \tag{8.52}$$

where $\vec{u} = \{u_{1_1}, u_{2_1}, \ldots, u_{i_{2^{i-1}}}\}^{\mathrm{T}}$ and $\boldsymbol{J}$ is a Jacobian matrix with elements $\partial_{ik} F_{i_j}/T_i$ evaluated at the fixed point $\boldsymbol{u} = \{1,1,\ldots,1\}$. We then find the eigenvalues by solving $\det(\boldsymbol{J} - \frac{1}{T}\mathbb{I}) = 0$ where $\mathbb{I}$ is the identity matrix. The characteristic polynomial of an $n \times n$ matrix can be expressed in terms of powers of the trace; however, roots of polynomials of degree five or more are unlikely to yield a closed-form solution in general.

8.4 Chapter summary

In this chapter we have considered the repeated percolation of random tree-like graphs extending the 2-strain epidemic models of Newman [49, 53]. We have discussed the outbreak size and the critical behaviour of the models. These experiments are similar to a repeated attack on a partially damaged network and so we have also plotted the degree distribution for each outbreak, for each topology and its cumulative value. We found that, for a coinfecting disease that spreads on the GCC of the preceding outbreaks, Poisson networks fracture more readily than scale-free networks; whilst the converse is true for cross-immune diseases that spread on the RG.

Chapter Nine

Conclusion

In this thesis, we have endeavoured to rationalise the properties of clustered configuration model networks using the analytically exact method generating functions. This topic is important because of the widespread usefulness of complex networks across multiple disciplines to describe networked interactions and the effect on the governing dynamics of the topology of those interactions. Analytical models tend to assume that the network is locally tree-like. Clustering could be defined as the failure to be tree-like; it is the tendency of contacts to aggregate into closed loops, inducing correlations into the locally tree-like assumption. There has been much research in the literature on the role that clustering plays on dynamical processes; in particular, bond percolation. It is often the case that changing the extent of clustering within the network also changes other properties, particularly within the configuration model, such as the degree assortativity, which is the tendency of high-degree contacts to preferentially connect together or not. Degree assortativity also plays a significant role on the percolation properties of a random graph; and so, clustering and assortativity are often studied together.

To inject clustering into a tree-like network, the generalised configuration model assumes that a vertex can belong to a number of predetermined motifs, such as cliques or other small cycles. The joint distribution of the vertices motif membership is then used to create random graphs with potentially tight clustering, upon which Monte Carlo simulation of percolation can be performed.

Analytically, the joint distribution can be used within the generating function method to provide an exact theoretical model (in the limit of infinitely sized networks) to complement the simulation.

In chapter 3, we have provided an exact closed-form expression for the percolation properties of cliques and chordless cycles via arguments from enumerative combinatorics. The expression is based on the probability that a vertex fails to be attached to the giant connected component despite its membership in a clique or chordless cycle. It was shown that this expression can be inverted to describe the probability that a vertex *does* belong to the giant component. This dual description was then exploited to introduce the complement problem, where the entire network was described, rather than either the giant component or the residual graph. It was then stated that this full description is the fundamental reasoning behind the utility of the partial immunity model we later discussed in chapter 7.

In chapter 4, we derived the distribution of finite components in random configuration model graphs that are composed of cliques. The finite components of single-topology

networks were studied and supported analytically, before being generalised to mixed-clique networks. There is more work to be done to understand the role of clustering on the distribution of components, their average size as a function of bond occupancy probability and the effect of assortativity on these results.

In chapter 5 we formalised the counting procedure for other motifs, which do not exhibit regular structure of cliques or chordless cycles and discussed the difficulties in finding an all-encompassing closed-form expression for their percolation properties. We then reviewed our approximate method, based on counting SIR trees rather than enumerating induced subgraphs.

In chapter 6, we derived the probability that motifs of different topologies are connected together and supported our expressions with Monte Carlo simulation. We then introduced a novel clique cover algorithm and showed that the ensemble of random graphs that can be created from our clique cover accurately describes the correlation properties of an empirical network, particularly the high-degree sites.

Chapter 7 defined a new avenue of research for the thesis to discuss epidemics on networks. In this chapter, we extended two important 2-stage epidemic models from the literature to the realm of networks with 3-cliques. We then generalised this interaction mechanism, based on the description of the giant and residual components from chapter 3 to a partial interaction model.

In chapter 8, in a break from all previous endeavours, we examined the repeated percolation of tree-like networks under the interaction conditions of cross-immunity and coinfection. We used this to probe the robustness of scale-free and Erdős-Renyi networks.

9.1 Future work

There many unfinished avenues of research from this work. These include the proper investigation of the properties of the finite components of random networks composed of cliques, chordless cycles or arbitrary motifs under percolation and controlled assortativity conditions.

I am intrigued to re-derive the approximate percolation model based on counting connected trees within motifs versus counting connected subgraphs. It would be interesting to see if I simply counted it incorrectly, and that it could be made exact. This would also hint at a deeper connection between the probabilities associated to percolation of trees and graphs.

I believe that the complement problem, where the entire network following percolation is fully described, will allow the analytical description of a Potts model, which could be likened to a contemporaneous multi-strain epidemic model. This would proceed via the principle of inclusion-exclusion, which is a counting technique in combinatorics to find the union of finite sets.

The correlation properties of networks should be investigated further. I think the current portrayal of zig-zag patterns is unintentionally misleading, and is mostly due to choice of neighbour on the x-axis (see Fig 6.2 for instance). It would be better to continue the analytical development of the model and introduce an overall degree-correlation function or similar. It would also be interesting to study this for bond percolation over a substrate network, or extend it to the repeated percolation epidemic model from chapter 8.

I think the properties of 2-strain epidemic on random graphs with cliques and chordless-cycles should be investigated. Most likely in the form of the partial immunity model.

Similarly, the N-strain epidemic model should be generalised to the case of clustered networks as well as the partial immunity interaction.

I would be very interested to see if an endemic equilibrium could be described in terms of N-strains of a partial immunity SIR model with fixed transmissibility.

Other studies should also be performed: such as addition-deletion processes (whereby motifs are added or deleted from random graphs); the effect of heterogeneous susceptibility and transmissibilities (percolation on semi-directed clustered networks); or the study of multilayer networks with multiple interaction pathogens, should be investigated also. I find the addition-deletion process very interesting, as this is essentially a birth-death process and it could potentially be coupled with an epidemic process also.

APPENDIX A

EQUIVALENT EXPRESSIONS FOR g_3^2

In this section we will examine some of the different expressions for g_3^2 that are currently in the literature, in a similar manner to the analysis of g_2 performed in section 2.6. Perhaps the earliest expression for g_3 in the network science literature is due to Newman in 2003 [47]; although earlier work in 1959 by Gilbert contained the same terms [17]. The expression is obtained from application of the recursive method for cliques, detailed in 3.2; however, it was not written out in full by either Gilbert or Newman. The polynomial that is obtained is also the one used by Karrer and Newman in 2010 [25]; Hasegawa *et al* [23] and is also given by Mann's Eq 3.21 and Eq 5.6 [36, 32]. It is written as

$$g_3^2 = (1-T)^2 + 2T(1-T)^2 u_3 + 3T^2(1-T)u^2 + T^3u^2 \tag{A.1}$$

The logic behind each term of this polynomial is graphically displayed in Fig 3.2 and is based on enumerating the ways that a focal vertex can remain attached to the RG.

In 2009 Newman extended the configuration model to account for random graphs with 2- and 3-clique subgraphs [52]. The expression that he used for g_3^2 is based on the inverse logic of section 3.3 and is given by

$$g_3^2 = 1 - 2T(1-T)^2 z_3 - 3T^2(1-T)(1-(1-z_3)^2) - T^3(1-(1-z_3)^2) \tag{A.2}$$

where $z_3 = 1-u_3$ and therefore $1-u_3^2 = (1-(1-z_3)^2)$. This expression is constructed as unity minus the number of ways that a vertex can be attached to the GCC, $g_3^2 = 1-f_3^2$. This is similar to the logic behind Eq A.1; however, the semantics of which graph (GCC or RG) we connect to are inverted.

Also in 2009, Miller contemporaneously extended the configuration model in the same manner as Newman [39]. Miller's logic is distinct from the first two counting methods

$$g_3^2 = (1-T+u_3T)^2 - 2u_3(1-u_3)T^2(1-T) \tag{A.3}$$

This expression first counts the probability that both edges connected to the focal vertex fail to connect it to the GCC as if they were independent of one another, which is simply $(g_2)^2$; minus the probability that the third edge is used to connect the focal vertex. For this

to occur, one of the neighbours must be attached to the GCC and fail to occupy its edge to the focal vertex; the other must initially be in the RG such that the first neighbour has a path around the cycle to reach the focal vertex. Miller's equation is the motivation behind Mann's approximate method to enumerate g_η for arbitrary cycles [32, 33].

The next expression for g_3^2 is due to Mann in 2021 [35]. In this case we assume that the focal vertex is a member of the RG and enumerates the ways in which its neighbours fail to attach it to the GCC as

$$g_3^2 = u_3^2 + [(1-u_3)(1-T)]^2 + 2u_3(1-u_3)(1-T)(1-T^2) \tag{A.4}$$

Specifically, from left to right, both neighbours can themselves belong to the GCC, both can be in the GCC but have unoccupied edges to the focal vertex or one can belong to the GCC (with the other in the RG) and both the 1-hop and 2-hop paths fail to be occupied.

The final expression we consider is also due to Mann in 2021 [31] and is based on the premise that the focal vertex is connected to the GCC. In the assumption, the embedded focal vertex has 3 types of neighbours: vertices in the RG whose full set of neighbours failed to occupy edges to connected it; vertices in the GCC that were attached by their neighbours other than the focal vertex; and finally, vertices in the RG whose neighbours have failed to connect it, but then the focal vertex *did* connect it to the GCC. Each pairwise combination of neighbour must then be enumerated and the resultant expression is given by

$$\begin{aligned} g_3^2 = {} & u_3^2(1-T_1)^2 + ((1-u_3)(1-T_1))^2 + (u_3T_1)^2 \\ & + 2u_3(1-T_1)^2(1-u_3)(1-T_1^2) + 2u_3(1-T_1)u_3T_1 \\ & + 2(1-u_3)u_3T_1(1-T_1)(1-T_1^2)) \end{aligned} \tag{A.5}$$

APPENDIX B

$q_{n,k}$

The number of connected graphs of n labelled vertices over k edges is given by $q_{n,k}$. This quantity has a well known recursion formula as well as a closed-form analytical solution [16, 67, 58, 47]. Given the importance of this quantity to the contents of this thesis, we will review this derivation now.

Let Q be the combinatorial class of connected graphs and G the combinatorial class of all labelled graphs [14]. The relation between these two classes is the set-of relation: a graph is a set of connected components. This indicates that the mixed exponential generating function $G(z,y)$ of G can be generated from $Q(z,y)$ according to the following relationship

$$G(z,y) = \exp Q(z,y) \tag{B.1}$$

For m vertices, there are a total of $\binom{m}{2} = m(m-1)/2$ potential edges not allowing self-loops or multi-edges between the vertices. This set has $2^{\binom{m}{2}}$ possible partitions. Therefore, counting vertices (with z) and edges (with y) we have that

$$G(z,y) = \sum_m \sum_l \frac{1}{m!} \binom{\binom{m}{2}}{l} z^m y^l \tag{B.2}$$

From the binomial theorem we find

$$G(z,y) = \sum_m \frac{z^m}{m!} (1+y)^{\binom{m}{2}} \tag{B.3}$$

or

$$G(z,y) = 1 + \sum_{m \geq 1} (1+y)^{m(m-1)/2} \frac{z^m}{m!} \tag{B.4}$$

This yields an expression for the entire series of connected labelled graphs, $Q(z,y)$; since, $Q(z,y) = \log G(z,y)$ such that we obtain

$$Q(z,y) = \log \left(1 + \sum_{m \geq 1} (1+y)^{m(m-1)/2} \frac{z^m}{m!} \right) \tag{B.5}$$

We can then perform a series expansion of the logarithm using

$$\log(1+x) = \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{k} x^k$$

to obtain

$$Q(z,y) = \sum_{l\geq 1} (-1)^{l+1} \frac{1}{l} \left(\sum_{m\geq 1} (1+y)^{m(m-1)/2} \frac{z^m}{m!} \right)^l \tag{B.6}$$

We now examine the case of n vertices and k edges where $k \geq n-1$ by extracting the coefficient $q_{n,k}$ of $[z^n][y^k]$. Note that the term in the parenthesis has minimum degree l in z, allowing us to disregard the series beyond $l > n$. This yields the formula for the number of connected labelled graphs with n vertices and k edges as

$$\begin{aligned} q_{n,k} = n![z^n][y^k] \sum_{l=1}^{n} (-1)^{l+1} \frac{1}{l} \\ \times \left(\sum_{m=1}^{n} (1+y)^{m(m-1)/2} \frac{z^m}{m!} \right)^l \end{aligned} \tag{B.7}$$

The coefficient of z^n is given by the integer partitions $\lambda \vdash n$ of length l, multiplied by their multiplicity (number of compositions)

$$\frac{1}{n!} \binom{n}{\lambda} \binom{l}{f} \tag{B.8}$$

where for partition λ we have $\lambda = 1^{f_1} 2^{f_2} 3^{f_3} \cdots$ and so on, such that we have

$$q_{n,k} = \sum_{\lambda \vdash n} \frac{(-1)^{l+1}}{l} \binom{n}{\lambda} \binom{l}{f} (1+y)^{\sum_{\lambda_i} \binom{\lambda_i}{2}} \tag{B.9}$$

for $\lambda_i \in \lambda$. The coefficient of y^k is found from the binomial theorem to yield a final expression for $q_{n,k}$ as

$$q_{n,k} = \sum_{\lambda \vdash n} \frac{(-1)^{l+1}}{l} \binom{n}{\lambda} \binom{l}{f} \binom{\sum_{\lambda_i} \lambda_i(\lambda_i - 1)/2}{k} \tag{B.10}$$

APPENDIX C

DEGREE CORRELATIONS WITHIN THE TREE-TRIANGLE MODEL

In this section we derive the expectation values for the tree-triangle model. For this model the generating function for the number of nearest-neighbours given the joint degree of the focal vertex is $k_{\boldsymbol{\tau},0} = (s_0, t_0)$ is given by unpacking Eq 6.17 for $\boldsymbol{\tau} = \{\perp, \Delta\}$. We obtain

$$\hat{F}_{GCC}(\boldsymbol{x}, \boldsymbol{y}, s_0, t_0) = p_{s_0,t_0} f_{\perp}^{s_0} f_{\Delta}^{2t_0} - p_{s_0,t_0} g_{\perp}^{s_0} g_{\Delta}^{2t_0} \tag{C.1}$$

where $f_\tau = \sum_s \sum_t q_{\tau,(s,t)} z_{st}$, $g_\perp = \sum_s \sum_t q_{\perp,(s,t)} u_\perp^{s-1} u_\Delta^{2t} x_s y_t$ and $\sum_s \sum_t q_{\Delta,(s,t)} u_\perp^{s} u_\Delta^{2(t-1)} x_s y_t$. The evaluation of the expectation values for the nearest-neighbours to a vertex of joint degree (s_0, t_0) in the tree-triangle model is given by the following derivative

$$\hat{F}'_{\mathrm{GCC}} = \left. \frac{d\hat{F}_{\mathrm{GCC}}}{dz_{s't'}} \right|_{z_{s't'}=1} \tag{C.2}$$

We evaluate this as follows

$$\left. \frac{d\hat{F}_{\mathrm{GCC}}}{dz_{s't'}} \right|_{z_{s't'}=1} = \left. \frac{d}{dz_{s't'}} \right|_{z_{s't'}=1} p_{s_0t_0} f_{\perp}^{s_0} f_{\Delta}^{2t_0} - \left. \frac{d}{dz_{s't'}} \right|_{z_{s't'}=1} p_{s_0t_0} g_{\perp}^{s_0} g_{\Delta}^{2t_0} \tag{C.3}$$

$$= p_{s_0t_0} \left(s_0 f_{\perp}^{s_0-1} \frac{df_{\perp}}{dz_{s't'}} f_{\Delta}^{2t_0} + 2t_0 f_{\perp}^{s_0} f_{\Delta}^{2(t_0-1)} f_{\Delta} \frac{df_{\Delta}}{dz_{s't'}} \right)$$

$$- p_{s_0t_0} \left(s_0 g_{\perp}^{s_0-1} \frac{dg_{\perp}}{dz_{s't'}} g_{\Delta}^{2t_0} + 2t_0 g_{\perp}^{s_0} g_{\Delta}^{2(t_0-1)} g_{\Delta} \frac{dg_{\Delta}}{dz_{s't'}} \right) \tag{C.4}$$

At $z_{s't'} = 1$ we have $f_\tau(1) = 1$, $g_\tau(1) = G_{1,\tau}(u_\perp, u_\Delta^2)$ and also

$$\left. \frac{df_\tau}{dz_{s't'}} \right|_{z_{s't'}=1} = \frac{d}{dz_{s't'}} \sum_s \sum_t q_{\tau,(s,t)} z_{st} \tag{C.5}$$

$$= q_{\tau,(s',t')} \tag{C.6}$$

and

$$\left.\frac{dg_\perp}{dz_{s't'}}\right|_{z_{s't'}=1} = \frac{d}{dz_{s't'}}\sum_s\sum_t q_{\perp,(s,t)}u_\perp^{s-1}u_\Delta^{2t}z_{st} \tag{C.7}$$

$$= q_{\perp,(s',t')}u_\perp^{s'-1}u_\Delta^{2t'} \tag{C.8}$$

$$\left.\frac{dg_\Delta}{dz_{s't'}}\right|_{z_{s't'}=1} = \frac{d}{dz_{s't'}}\sum_s\sum_t q_{\Delta,(s,t)}u_\perp^{s}u_\Delta^{2(t-1)}z_{st} \tag{C.9}$$

$$= q_{\Delta,(s',t')}u_\perp^{s'}u_\Delta^{2(t'-1)} \tag{C.10}$$

Thus, we find

$$\left.\frac{d\hat{F}_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1} = p_{s_0t_0}\left(s_0q_{\perp,(s',t')}+2t_0q_{\Delta,(s',t')}\right) - p_{s_0t_0}\left(s_0u_\perp^{s_0-1}q_{\perp,(s',t')}u_\perp^{s'-1}u_\Delta^{2t'}u_\Delta^{2t_0}\right.$$
$$\left.+2t_0u_\perp^{s_0}u_\Delta^{2(t_0-1)}u_\Delta q_{\Delta,(s',t')}u_\perp^{s'}u_\Delta^{2(t'-1)}\right) \tag{C.11}$$

The evaluation of the expectation values for the nearest-neighbours to the average vertex in the tree-triangle model is given by the following derivative

$$F'_{\mathrm{GCC}} = \sum_{s'}\sum_{t'}\left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1} \tag{C.12}$$

where F_{GCC} is given by unpacking Eq 6.20 for $\boldsymbol{\tau}=\{\perp,\Delta\}$ to find

$$F_{GCC}(\boldsymbol{x},\boldsymbol{y}) = \sum_s\sum_t p_{s,t}f_\perp^s f_\Delta^{2t} - \sum_s\sum_t p_{s,t}g_\perp^s g_\Delta^{2t} \tag{C.13}$$

To evaluate this consider the following derivative

$$\left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1} = \left.\frac{d}{dz_{s't'}}\right|_{z_{s't'}=1}G_0(f_\perp,f_\Delta) - \left.\frac{d}{dz_{s't'}}\right|_{z_{s't'}=1}G_0(g_\perp,g_\Delta) \tag{C.14}$$

$$= \left.\frac{d}{dz_{s't'}}\right|_{z_{s't'}=1}\sum_s\sum_t p_{st}f_\perp^s f_\Delta^{2t} - \left.\frac{d}{dz_{s't'}}\right|_{z_{s't'}=1}\sum_s\sum_t p_{st}g_\perp^s g_\Delta^{2t} \tag{C.15}$$

$$= \sum_s\sum_t p_{st}\left\{sf_\perp^{s-1}\frac{df_\perp}{dz_{s't'}}f_\Delta^{2t}+2tf_\perp^s f_\Delta^{2(t-1)}f_\Delta\frac{df_\Delta}{dz_{s't'}}\right\}$$
$$-\sum_s\sum_t p_{st}\left\{sg_\perp^{s-1}\frac{dg_\perp}{dz_{s't'}}g_\Delta^{2t}+2tg_\perp^s g_\Delta^{2(t-1)}g_\Delta\frac{dg_\Delta}{dz_{s't'}}\right\} \tag{C.16}$$

When evaluated at $z_{(s',t')}=1$ we have that $f_\tau(1)=1$ and so the first bracket simplifies significantly. The second bracket is more involved; however, using the self-consistent expressions for $u_\perp = G_{1,\perp}(u_\perp,u_\Delta^2)$ and $u_\Delta = G_{1,\Delta}(u_\perp,u_\Delta^2)$ we can write $g_\perp(1)=u_\perp$ and

$g_\Delta(1) = u_\Delta$ to obtain

$$\left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1} = \sum_s \sum_t p_{st}\left\{ sq_{\perp,(s',t')} + 2tq_{\Delta,(s',t')} \right\} - \sum_s \sum_t p_{st}\left\{ su_\perp^{s-1} q_{\perp,(s',t')} u_\perp^{s'-1} u_\Delta^{2t'} u_\Delta^{2t} + 2tu_\perp^s u_\Delta^{2(t-1)} u_\Delta q_{\Delta,(s',t')} u_\perp^{s'} u_\Delta^{2(t'-1)} \right\} \tag{C.17}$$

We now sum over (s',t') to obtain

$$\begin{aligned}\sum_{s'}\sum_{t'} \left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1} &= \sum_s \sum_t p_{st}\left\{ s\sum_{s'}\sum_{t'} q_{\perp,(s',t')} + 2t\sum_{s'}\sum_{t'} q_{\Delta,(s',t')} \right\} \\ &\quad - \sum_s \sum_t p_{st}\left\{ su_\perp^{s-1} u_\Delta^{2t} \sum_{s'}\sum_{t'} q_{\perp,(s',t')} u_\perp^{s'-1} u_\Delta^{2t'} \right. \\ &\quad \left. + 2tu_\perp^s u_\Delta^{2(t-1)} u_\Delta \sum_{s'}\sum_{t'} q_{\Delta,(s',t')} u_\perp^{s'} u_\Delta^{2(t'-1)} \right\}\end{aligned} \tag{C.18}$$

The probability distributions are normalised and hence have the following property $\sum_s \sum_t q_{\tau,(s,t)} = 1$, so the first bracket reduces trivially to the sum of the average degrees of each edge topology. The second bracket also reduces; dealing first with the double summation over dashed variables we find

$$\sum_{s'}\sum_{t'} \left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1} = \sum_s \sum_t p_{st}\,(s+2t) - \sum_s \sum_t p_{st}\left\{ su_\perp^{s-1} u_\Delta^{2t} u_\perp + 2tu_\perp^s u_\Delta^{2(t-1)} u_\Delta^2 \right\} \tag{C.19}$$

before observing that

$$\sum_s \sum_t p_{st} s x^{s-1} y^t = \langle s \rangle G_{1,\perp}(x,y) \tag{C.20}$$

$$\sum_s \sum_t p_{st} t x^s y^{t-1} = \langle t \rangle G_{1,\Delta}(x,y) \tag{C.21}$$

to arrive at

$$\sum_{s'}\sum_{t'} \left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1} = \langle s \rangle + 2\langle t \rangle - \langle s \rangle G_{1,\perp}(u_\perp, u_\Delta^2) u_\perp - 2\langle t \rangle G_{1,\Delta}(u_\perp, u_\Delta^2) u_\Delta^2 \tag{C.22}$$

Substituting the self-consistent relationships for $u_\perp$ and u_Δ we finalise the expression as

$$\sum_{s'}\sum_{t'} \left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1} = \langle s \rangle \left(1 - u_\perp^2\right) + 2\langle t \rangle \left(1 - u_\Delta^3\right) \tag{C.23}$$

In the case that there are no triangles present in the model, then $u_\Delta = 1$ and $\langle t \rangle = 0$; the expression reduces to

$$\sum_{s'} \left.\frac{dF_{\mathrm{GCC}}}{dz_{s'}}\right|_{z_{s'}=1} = \langle s \rangle \left(1 - u_\perp^2\right) \tag{C.24}$$

which is the result of [41] in the case that $l = 1$. In the opposite case, when there are no ordinary edges, we find

$$\sum_{t'} \frac{dF_{\mathrm{GCC}}}{dz_{t'}}\bigg|_{z_{t'}=1} = 2\langle t\rangle \left(1-u_{\Delta}^{3}\right) \tag{C.25}$$

The probability $P(k_{\tau,0},k_{\tau,1}) = P((s_0,t_0),(s',t'))$ is given by the quotient of Eqs C.11 and C.23 where we find

$$\begin{aligned} P((s_0,t_0),(s',t')) &= \frac{d\hat{F}_{\mathrm{GCC}}}{dz_{s't'}}\bigg|_{z_{s't'}=1} \Bigg/ \sum_{s'}\sum_{t'} \frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\bigg|_{z_{s't'}=1} \\ &= p_{s_0t_0}\left(s_0 q_{\perp,(s',t')} + 2t_0 q_{\Delta,(s',t')}\right) - p_{s_0t_0}\Bigg(s_0 u_{\perp}^{s_0-1} q_{\perp,(s',t')} u_{\perp}^{s'-1} u_{\Delta}^{2t'} u_{\Delta}^{2t_0} \\ &\quad + 2t_0 u_{\perp}^{s_0} u_{\Delta}^{2(t_0-1)} u_{\Delta} q_{\Delta,(s',t')} u_{\perp}^{s'} u_{\Delta}^{2(t'-1)}\Bigg) \Bigg/ \langle s\rangle\left(1-u_{\perp}^{2}\right) + 2\langle t\rangle\left(1-u_{\Delta}^{3}\right) \end{aligned} \tag{C.26}$$

The conditional probability that a neighbour has joint degree (s',t') given a focal vertex of joint degree (s_0,t_0) is

$$P(s',t' \mid s_0,t_0) = \frac{p_{s_0t_0} s_0 q_{\perp,(s',t')}[1-u_{\perp}^{s_0+s'-2}u_{\Delta}^{2(t_0+t')}] + 2p_{s_0t_0}t_0 q_{\Delta,(s',t')}[1-u_{\perp}^{s_0+s'}u_{\Delta}^{2(t_0+t'-2)+1}]}{p_{s_0t_0}(s_0+2t_0)[1-u_{\perp}^{s_0}u_{\Delta}^{2t_0}]} \tag{C.27}$$

Using Eq 6.29 we find the average joint degree of a neighbour to a (s_0,t_0) vertex as

$$\mathscr{E}[\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}] = \left(\sum_{s',t'} s'P(s',t' \mid s_0,t_0), \sum_{s',t'} t'P(s',t' \mid s_0,t_0)\right)^{T} \tag{C.28}$$

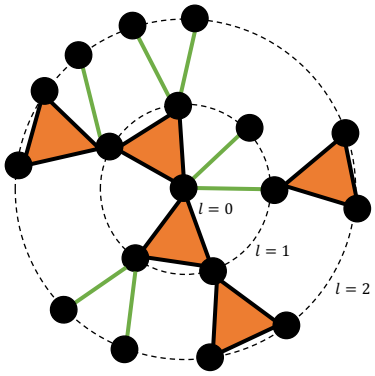


Figure C.1: An example of the degree correlation model in the tree-triangle model; 3-cliques are shaded orange whilst 2-cliques are coloured green. The joint degree of the focal vertex in layer $l = 0$ is $k_{\boldsymbol{\tau},0} = (2,2)$. We can follow edges of topology $\perp$ or Δ to the first neighbours. The distribution of the joint degrees of vertices in layer $l = 2$ depends on the topology of the path that we choose to reach it. Note, we do not traverse edges between triangles that lead to vertices in the same layer.

BIBLIOGRAPHY

[1] Aristoteles and David Ross. *The Nicomachean ethics of Aristotle*. Oxford University Press, 1980.

[2] Piotr Bialas and Andrzej K. Oleś. Correlations in connected random graphs. *Phys. Rev. E*, 77:036124, Mar 2008.

[3] Marián Boguñá and Romualdo Pastor-Satorras. Epidemic spreading in correlated complex networks. *Phys. Rev. E*, 66:047104, Oct 2002.

[4] Marián Boguñá, Romualdo Pastor-Satorras, and Alessandro Vespignani. Absence of epidemic threshold in scale-free networks with degree correlations. *Phys. Rev. Lett.*, 90:028701, Jan 2003.

[5] Tom Britton, Maria Deijfen, Andreas N. Lageras, and Mathias Lindholm. Epidemics on random graphs with tunable clustering. *Journal of Applied Probability*, 45(3):743–756, 2008.

[6] Giulio Burgio, Alex Arenas, Sergio Gómez, and Joan T. Matamalas. Network clique cover approximation to analyze complex contagions through group interactions. *arXiv e-prints*, page arXiv:2101.03618, January 2021.

[7] Duncan S. Callaway, M. E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85(25):5468–5471, 2000.

[8] Reuven Cohen, Keren Erez, Daniel Ben-Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Physical Review Letters*, 85(21):4626–4628, 2000.

[9] Pol Colomer-de Simón and Marián Boguñá. Double percolation phase transition in clustered complex networks. *Phys. Rev. X*, 4:041020, Oct 2014.

[10] Simon Dobson. *Epidemic modelling: Some notes, Maths, and code*. Independent Publishing Network, 2020.

[11] K. T. D. Eames. Modelling disease spread through random and regular contacts in clustered populations. *Theoretical Population Biology*, 73(1):104–111, 2008.

[12] Víctor M. Eguíluz and Konstantin Klemm. Epidemic threshold in structured scale-free networks. *Phys. Rev. Lett.*, 89:108701, Aug 2002.

[13] Michael E. Fisher and John W. Essam. Some cluster size and percolation problems. *Journal of Mathematical Physics*, 2(4):609–619, 1961.

[14] Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics*. Cambridge University Press, 2013.

[15] Ira M. Gessel. Lagrange inversion. *Journal of Combinatorial Theory, Series A*, 144:212–249, 2016.

[16] E. N. Gilbert. Enumeration of labelled graphs. *Canadian Journal of Mathematics*, 8:405–411, 1956.

[17] E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[18] James P. Gleeson. Bond percolation on a class of clustered random networks. *Physical Review E*, 80(3), Oct 2009.

[19] James P. Gleeson and Sergey Melnik. Analytical results for bond percolation and k-core sizes on clustered networks. *Phys. Rev. E*, 80:046121, Oct 2009.

[20] James P. Gleeson, Sergey Melnik, and Adam Hackett. How clustering affects the bond percolation threshold in complex networks. *Phys. Rev. E*, 81:066114, Jun 2010.

[21] P. Grassberger. On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, 63(2):157–172, 1983.

[22] Adam Hackett, Sergey Melnik, and James P. Gleeson. Cascades on a class of clustered random networks. *Phys. Rev. E*, 83:056107, May 2011.

[23] Takehisa Hasegawa and Shogo Mizutaka. Structure of percolating clusters in random clustered networks. *Phys. Rev. E*, 101:062310, Jun 2020.

[24] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30:175–193, 1906.

[25] Brian Karrer and M. E. J. Newman. Random graphs containing arbitrary distributions of subgraphs. *Physical Review E*, 82(6), 2010.

[26] Brian Karrer and M. E. J. Newman. Competing epidemics on complex networks. *Phys. Rev. E*, 84:036106, Sep 2011.

[27] Eben Kenah and James M. Robins. Second look at the spread of epidemics on networks. *Phys. Rev. E*, 76:036113, Sep 2007.

[28] Istvan Z. Kiss and Darren M. Green. Comment on "properties of highly clustered networks". *Phys. Rev. E*, 78:048101, Oct 2008.

[29] Ivan Kryven. Finite connected components in infinite directed and multiplex networks with arbitrary degree distributions. *Phys. Rev. E*, 96:052304, Nov 2017.

[30] Elizabeth Leicht. Percolation on interacting networks. In *International Workshop and Conference on Network Science (NetSci)*, 2009.

[31] Peter Mann, V. Anne Smith, John B. O. Mitchell, and Simon Dobson. Cooperative coinfection dynamics on clustered networks. *Phys. Rev. E*, 103:042307, Apr 2021.

[32] Peter Mann, V. Anne Smith, John B. O. Mitchell, and Simon Dobson. Percolation in random graphs with higher-order clustering. *Phys. Rev. E*, 103:012313, Jan 2021.

[33] Peter Mann, V. Anne Smith, John B. O. Mitchell, and Simon Dobson. Random graphs with arbitrary clustering and their applications. *Phys. Rev. E*, 103:012309, Jan 2021.

[34] Peter Mann, V. Anne Smith, John B. O. Mitchell, and Simon Dobson. Symbiotic and antagonistic disease dynamics on networks using bond percolation. *Phys. Rev. E*, 104:024303, Aug 2021.

[35] Peter Mann, V. Anne Smith, John B. O. Mitchell, and Simon Dobson. Two-pathogen model with competition on clustered networks. *Phys. Rev. E*, 103:062308, Jun 2021.

[36] Peter Mann, V. Anne Smith, John B. O. Mitchell, Christopher Jefferson, and Simon Dobson. Exact formula for bond percolation on cliques. *Phys. Rev. E*, 104:024304, Aug 2021.

[37] Sergey Melnik, Adam Hackett, Mason A. Porter, Peter J. Mucha, and James P. Gleeson. The unreasonable effectiveness of tree-based theory for networks with clustering. *Phys. Rev. E*, 83:036112, Mar 2011.

[38] Joel C. Miller. Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. *Phys. Rev. E*, 76:010101, Jul 2007.

[39] Joel C. Miller. Percolation and epidemics in random clustered networks. *Phys. Rev. E*, 80:020901, Aug 2009.

[40] Joel C Miller. Spread of infectious disease through clustered populations, Dec 2009.

[41] Shogo Mizutaka and Takehisa Hasegawa. Emergence of long-range correlations in random networks. *Journal of Physics: Complexity*, 1(3):035007, sep 2020.

[42] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.

[43] Cristopher Moore and M. E. J. Newman. Exact solution of site and bond percolation on small-world networks. *Phys. Rev. E*, 62:7059–7064, Nov 2000.

[44] Moreno, Y. and Vázquez, A. Disease spreading in structured scale-free networks. *Eur. Phys. J. B*, 31(2):265–271, 2003.

[45] M. E. J. Newman. Spread of epidemic disease on networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 66 1 Pt 2:016128, 2002.

[46] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66:016128, Jul 2002.

[47] M. E. J. Newman. Properties of highly clustered networks. *Phys. Rev. E*, 68:026121, Aug 2003.

[48] M. E. J. Newman. Power laws, pareto distributions and Zipfs law. *Contemporary Physics*, 46(5):323–351, 2005.

[49] M. E. J. Newman. Threshold effects for two pathogens spreading on a network. *Physical Review Letters*, 95(10), Feb 2005.

[50] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, Sep 2006.

[51] M. E. J. Newman. Component sizes in networks with arbitrary degree distributions. *Phys. Rev. E*, 76:045101, Oct 2007.

[52] M. E. J. Newman. Random graphs with clustering. *Phys. Rev. Lett.*, 103:058701, Jul 2009.

[53] M. E. J. Newman and Carrie R. Ferrario. Interacting epidemics and coinfection on contact networks. *PLoS ONE*, 8(8), 2013.

[54] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2), 2001.

[55] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, Jul 2001.

[56] M. E.J. Newman. *Networks*. Oxford University Press, 2019.

[57] Romualdo Pastor-Satorras, Alexei Vázquez, and Alessandro Vespignani. Dynamical and correlation properties of the internet. *Phys. Rev. Lett.*, 87:258701, Nov 2001.

[58] R. J. Riddell and G. E. Uhlenbeck. On the theory of the virial development of the equation of state of monoatomic gases. *The Journal of Chemical Physics*, 21(11):2056–2064, 1953.

[59] Martin Ritchie, Luc Berthouze, and Istvan Z. Kiss. Generation and analysis of networks with a prescribed degree sequence and subgraph family: higher-order structure matters. *Journal of Complex Networks*, 5(1):1–31, 05 2016.

[60] Alejandro F. Rozenfeld, Reuven Cohen, Daniel ben Avraham, and Shlomo Havlin. Scale-free networks on lattices. *Phys. Rev. Lett.*, 89:218701, Nov 2002.

[61] L.M. Sander, C.P. Warren, I.M. Sokolov, C. Simon, and J. Koopman. Percolation on heterogeneous networks as a model for epidemics. *Mathematical Biosciences*, 180(1-2):293–305, 2002.

[62] M. Ángeles Serrano and Marián Boguñá. Clustering in complex networks. i. general formalism. *Phys. Rev. E*, 74:056114, Nov 2006.

[63] M. Ángeles Serrano and Marián Boguñá. Clustering in complex networks. ii. percolation properties. *Phys. Rev. E*, 74:056115, Nov 2006.

[64] Chaoming Song, Shlomo Havlin, and Hernán A. Makse. Self-similarity of complex networks. *Nature*, 433(7024):392–395, 2005.

[65] Ido Tishby, Ofer Biham, Eytan Katzav, and Reimer Kühn. Revealing the microstructure of the giant component in random graph ensembles. *Phys. Rev. E*, 97:042318, Apr 2018.

[66] C. P. Warren, L. M. Sander, and I. M. Sokolov. Firewalls, Disorder, and Percolation in Epidemics. *arXiv e-prints*, pages cond–mat/0106450, June 2001.

[67] Herbert S. Wilf. *Generatingfunctionology*. Academic Press, 1994.